POLITECNICO DI TORINO

Department of Management and Production Engineering

Master of Science in

Engineering and Management

Prediction of the Dominant Instruction Set Architecture in Data Centers

Supervisor

Prof. Marco Cantamessa

Candidate

Mahammad Latifli 274338

Academic Year 2021-2022

Table of Contents

Introduction

In the age of technological improvements theory of Dominant Designs has been a topic of high interest and research during the last decades. As a completely new design or colossal upgrade found and implemented in any field may even disrupt the whole market dominance of the firms, for the investors and companies it is very hard to predict potential dominant designs from their early stages to adapt their strategy for the potential threat to convert it to an opportunity instead and thus, also contribute to the development of the right technology. This research aims on developing a methodology for predicting the potential dominant design for the near future in Data Centers of the World focusing on the logic board architectures that are currently being implemented and being developed.

The first chapter of the thesis gives solid background information about the main research done in the field of dominant designs field so far including the outstanding work of the pioneers of the concept - Abernathy and Utterback. The evolution of the concept is also described according to the different perspectives and researchers.

The next chapter will offer background information on current computers and a comprehensive technical comparison of two popular computer architectures: Reduced Instruction Set Computer (RISC) and Complex Instruction Set Computer (CISC). CISC architecture which is currently a dominant design in Data Centers is being challenged by RISC architecture which seems very promising.

The last chapter focuses on establishing the mentioned Methodology to predict the future of the industry from the beginning stages. To build the methodology integrative framework of Fernando F.Suarez (2003) was used as a reference point. The methodology focuses on decision-making based on analysis of main variables that may affect the dominance of designs in the market at the industry level, rather than the firm level. The built framework can be very helpful for potential investors and firms for making key decisions by analyzing what the future would bring for the design, which barriers there are, and which of the factors need more resources and attention to achieve success following the selected design. This framework can also be handy for the already established competing firms to evaluate the potential threat for them in the industry. As a result of the analysis in the case of the battle between RISC and CISC, it is concluded that there is a very high chance for RISC to become a dominant design in the Global Data Centers market being superior in almost all the variables of the framework.

1. Dominant Design Concept

1.1. Evolution of Dominant Design concept

The term "Dominant Design" initially appeared in the literature in 1975 by William J. Abernathy and James M. Utterback, and it quickly became a focus of interest for several research institutes. It stems from the idea that in most new product categories, the market chooses a single product's architecture design as the primary one that defines the category's characteristics. The establishment of a dominant design is a required process that impacts the strategy and performance of enterprises developing that product dramatically. Additionally, a dominant design influences future generations of products in a certain category, culminating in the creation of an "architectural franchise" for the firm that created the dominant design and potentially excluding competitors (Schilling 1998). At the market's early stages of development, technical uncertainty and a large market result in a diversity of product designs (Abernathy and Utterback 1978). However, history shows that dominant design is not always the one with the best technical performance metrics; rather, it is the design that maximizes technological potential because of the alignment of interests among suppliers, users, and rivals.

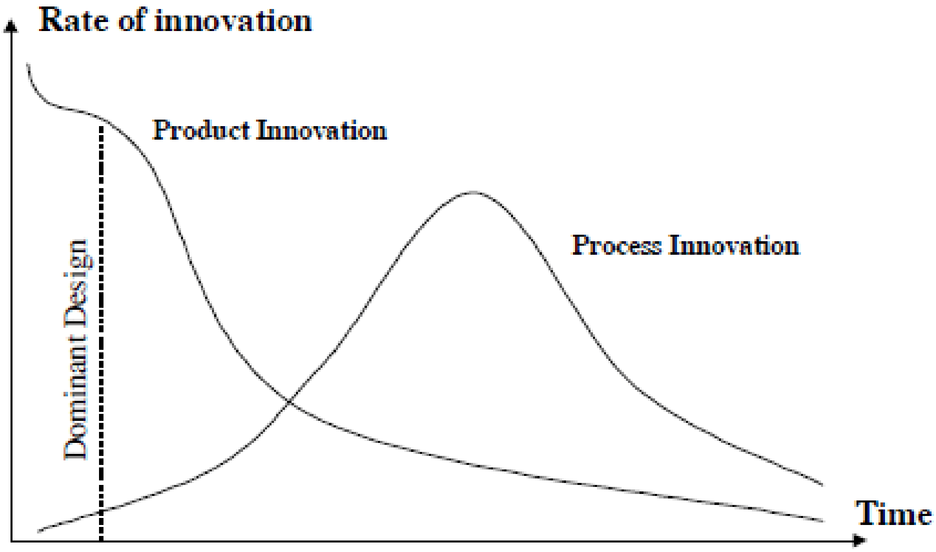
The concept's fundamental premise concerning the development of the proposed production process is that it evolves toward an increasing level of output productivity: capital intensity increases, productivity increases through specialization and a more efficient division of labor, product design becomes more standardized, and material flow within the process becomes more linear. Indeed, the model is limited to the assembly business, eliminating the process or service industries. The progress of technology and the economic system necessarily casts doubt on this model from a variety of perspectives, failing to account for numerous aspects that cannot be neglected for the present and future. The unit of analysis is not often the business, but rather the complete manufacturing process used to manufacture a product. The model's primary premise is that production processes tend to evolve and alter in a consistent and identifiable manner over time. Additionally, the model has been classified by three development phases that are consistent across all sectors studied and may be identified based on the characteristics of the production factors:

- **Fluid Phase:** Both the product and the process undergo rapid change, and the competitive landscape is highly diverse. The process is defined as "fluid" at this phase, with unstable connections between the various process elements characterized mostly by custom and non-standardized procedures. The production system is easily adaptable to changes in the sector during this period, but it is inefficient. Equipment is usually general-purpose and needs highly skilled labor. Inputs are limited to generally available resources.

- **Transitional Phase:** The manufacturing method matures, the product matures, and price rivalry intensifies. Due to the integration of automation and precise and strict process controls, operations become more specialized, business routines develop, and the production system becomes more efficient. Certain threads can be highly automated, while others remain largely manual, resulting in a rather "segmented" manufacturing process. Equipment is usually automated in sub-processes. Specialized inputs may be needed for some necessary elements.

- **Specific Phase:** Investments in the business increase significantly, and the development of the manufacturing process achieves an extremely high degree of efficiency, to the point where process improvements become increasingly impossible. Because the process is so closely linked, any adjustments are extremely costly, as even little changes affect other components of the process or the product's design. Equipment is generally fully automated needing only supervision and maintenance.

The Abernathy-Utterback (A-U) model has had a substantial impact on innovation studies and has been adopted by many scholars. Although most studies cite Abernathy and Utterback 1978, the model does not depict the main design concept. As a result, the model utilized by Abernathy and Utterback 1978 differs from the A-U model that is mainly referred to. Numerous researchers have accepted the A-U models which were developed through the combination of three papers which are Utterback and Abernathy 1975, 1978, and Abernathy 1978. Abernathy 1978 concluded the A-U model which is the model that is imagined. By evaluating the A-U model formation process through the three significant publications stated above, it is obvious that the definitive model is the one in Abernathy 1978.



Fluid Normal direction of development transition → Specific

Figure 1. A-U model in Abernathy (1978)

From *"Productivity Dilemma: Roadblock to innovation in the automobile Industry",* by W. J. Abernathy, 1978, Baltimore

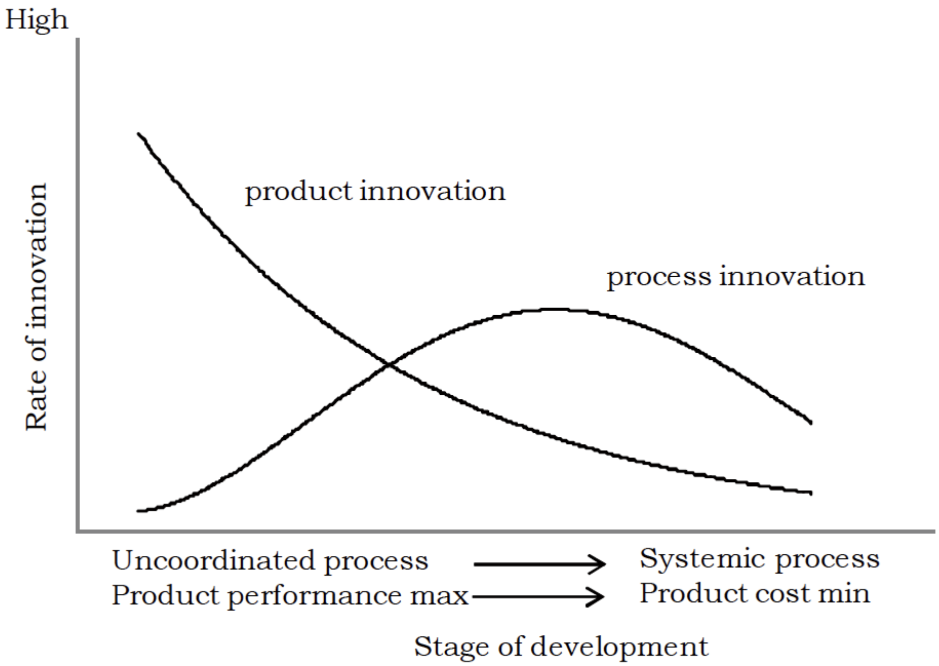


Figure 2. A-U model in Utterback and Abernathy (1975)

From *"A dynamic model of process and product innovation"*, by
J. M. Utterback and W. J. Abernathy, 1975, Omega

The rate of innovation during the development phase according to the three phases of development of the products is represented in Fig 1. The dotted perpendicular line represents the release of the Dominant Design. As observed on the graph, rate of the product innovation drops while the industry matures as there is more focus on process innovation to make the production and development process of the product more efficient. At the specific phase, however, both process and product innovation tend to drop as at that stage driving overall efficiency to increase becomes a more challenging task.

1.2. Process Development Model

A manufacturing process is the collection of process equipment, labor, job specifications, material inputs, work, and information flows used to create a product or service. The basic premise of the proposed model of process development is that as a production process evolves toward higher levels of output productivity over time, it follows a distinct evolutionary pattern: it becomes more highly capital-intensive, direct labor productivity increases through increased specialization of tasks, the flow of materials within the process takes on a more straight-line flow quality, and the process itself becomes more efficient. Increasing the productivity of a process by making incremental modifications to these parameters has a compounding effect, which improves the process in profound ways. The pattern of changes between stages in the process is pervasive, extending beyond the physical characteristics to the productivity variables themselves. There may also be changes in the internal organizational structure, the creation of a supply industry for certain materials, and technology-based capital goods as a process develop. The Abernathy and Utterback model classifies process development into three distinct stages which are uncoordinated, segmental, and systemic.

- **Uncoordinated:** During the early stages of a process or product's existence, market expansion and redefinition frequently result in competitive advances. Product and process innovations occur at a rapid pace, and competitors provide a wide variety of products. Typically, the process is constituted primarily of unstandardized and manual procedures, or operations involving general-purpose equipment. The process is fluid at this condition, with loose and unsettled interactions between process parts. This type of system is "organic" and adapts readily to environmental changes, but it is inevitably "slack" and "inefficient."

- **Segmental:** Price competition grows more intense as an industry and its product group develop. As production systems become more efficient, they become mechanistic and rigid. As tasks grow more specialized and subject to more rigorous operating rules, they become more formalized. Process-wise, the manufacturing system tends to grow more complicated and tightly integrated as a result of automation and process control. Certain subprocesses may be extensively automated using process-specific technology, while others may remain largely manual or rely on general-purpose equipment. As a result, in this scenario, industrial processes will have a fragmented quality. However, such significant development cannot proceed until a product group has matured sufficiently to generate enough sales and at least a few reliable product designs.

- **Systemic:** As a process grows more established and integrated, and an investment in it increases, it becomes increasingly difficult to improve individual process aspects selectively. Because the process becomes so closely integrated, adjustments become prohibitively expensive, as even slight changes may necessitate changes to other process parts and product design. At this level, process redesign is normally more gradual, but it may be prompted by the emergence of new technology or by a sudden or cumulative shift in market demand. If resistance to change persists as process technology and the market evolves, the stage is prepared for economic deterioration or revolutionary rather than evolutionary transformation.

The underlying concept is that a process, or productive segment, tends to evolve and alter in a predictable and identifiable manner over time.

1.3. Product Development Model

Product innovation is the commercialization of new technology or a mix of technologies to address a market segment's needs. Abernathy and Utterback's model proposes that products are developed predictably over time with an initial strategic emphasis on product performance (Performance maximization), followed by a focus on product variety (Sales maximization), and finally on standardization and cost reduction (Cost minimization):

- **Performance-Maximization:** Introduction of a technologically advanced product with a stronger emphasis on the product's uniqueness and performance. The sector is most often constituted of a few enterprises, new and small or longer-lived, that enter a new market by using their technological skills. At both the product and process levels, which correspond to the fluid phase, the market is poorly defined, the products lack uniformity, and the manufacturing method is primitive. Innovation is frequently motivated or inspired by new market needs (or possibilities), and its efficiency requires careful identification of product requirements rather than performance enhancements based on new scientific findings or even more advanced technology.

- **Sales-Maximization:** Both manufacturers and end consumers are assumed to have some knowledge of technology and product, significantly reducing market uncertainty, increasing competition primarily based on differentiation, and allowing for the emergence of some product designs as a sector standard. As a result, a large variety of products or the introduction of new components is possible. This phase of product innovation is comparable to the transition phase of process development. Process level changes can be triggered by a significant rise in output demand, resulting in relatively discontinuous process innovation that necessitates a new organizational structure in response to changes in production or product design.

- **Cost-Minimization:** During this phase, the market is firmly defined, and the product's life cycle evolves, becoming more standardized and lowering the product's diversity. Competition occurs primarily at the price level during this phase, eroding profit margins and limiting the number of enterprises operating in the industry, which effectively becomes an oligopoly in which efficiency and economies of scale become paramount. Thus, production becomes more "capital intensive," with the primary goal of lowering the costs of production inputs. Each change at this stage entails significant interdependent changes to the product and process, with extremely high costs and only limited advantages. Above all, innovation occurs at the level of suppliers, who benefit from far more favorable incentives and can be adopted by large enterprises functioning in the field.

1.4. Evolution of the Definition of Dominant Design

The definition of the dominant designs has changed from being broad and effects oriented to more specific during its evolution. Table 1.1 demonstrates the evolution of definitions of dominant design during its lifetime in the literature.

Source	**Definition of "Dominant Design"**
Abernathy and Utterback (1978)	A dominant design is a single architecture that establishes dominance in a product category.
Anderson and Tushman (1990)	A dominant design is a single architecture that establishes dominance in a product category.
Utterback (1994)	The design that is dominant in a product category is the one that commands the marketplace's allegiance. It is the standard to which competitors and innovators must comply if they wish to command a sizable market share. A dominating design is a product in a product category that achieves widespread acceptability as the technological standard for other market competitors to follow to earn considerable market share.
Suaréz and Utterback (1995)	The dominant design is a specific path along with an industry's design hierarchy that establishes dominance among competing design paths.
Christensen, Suaréz, and Utterback (1998)	A dominant design emerges in a product category when one product's design specifications (consisting of a single or a complement of design features) define the product category's architecture.

Table 1.1 Alternative definitions of "Dominant Design" in the literature

1.5. Dominant Designs versus Standards

In some prior research dominant designs and standards were used interchangeably which concludes these concepts weren't distinguished by the researchers (e.g., Anderson and Tushman 1990; Besen and Farrell 1994; Katz and Shapiro 1986; Schilling 1998). However, these two concepts are completely dissimilar. This article provides differences between standards and dominant designs as it is vital to understand how to distinguish them. Nowadays the term "standards" is widely used in engineering disciplines to define technical specifications such as quality, adaptability, connectivity to achieve proper functioning and user interaction, the user experience of designed products. Standards are a vital requirement for most products as the products themselves (e.g., wireless earbuds and phone) or their components (GPU and PC) should connect. Thus, standards in products serve a functional purpose and don't depend on market acceptance. However, market acceptance is an aspect of dominant designs.

1.6. Factors affecting technology dominance

1.6.1 Dominance Process

The technological domination of firms competing in the same field can be defined by several milestones that they pass during their evolution (F. Suarez, 2003). The technological field begins when a pioneering corporation or research group conducts applied R&D intending to develop a product, which can be considered as the first milestone. A second milestone is the emergence of the new product's first functional prototype. The first real prototype provides a strong signal to all competitors that at least one of the technology paths is possible and has been developed to the point where a product will soon be available. A working prototype frequently serves as a signal to competing firms that their research plans are feasible. A third milestone in the process of dominance is the launch of the first commercial product that establishes a direct link between laboratory technology and customers for the first time. Typically, the initial product on the market is prohibitively expensive for the mass market and is thus targeted at the upper end of the market. Although the early market is often modest in comparison to the mass market, it helps a particular design become an early flagship. The existence of a clear front-runner is the fourth and final milestone in the battle for dominance. Indeed, the forerunner has a possibility of prevailing, as its larger installed base tends to generate a preference for the technology with the highest market share. The result will be determined by how quickly competitors improve their solutions and how quickly the market expands. Katz and Shapiro's (1985) research demonstrates that when markets grow rapidly, initial excess inertia can be overcome by competing firms securing enough market share to rapidly expand their own installed bases, particularly if their product is superior to the flagship. Finally, at some point, a certain technological trajectory obtains supremacy, completing the dominance process. Fig. 3 depicts the timeline reported by the different milestones in the technology dominance process, where:

- t0 - the emergence of the technology and start of R&D.
- tP - the emergence of the new product prototype.
- tL - launch of the commercial product.
- tF - the appearance of the front-runner.
- tD - one of the alternative designs becomes dominant.

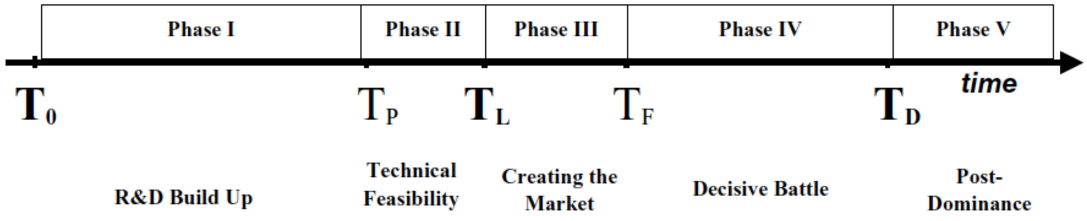


Fig. 3 Five milestones in the process of emergence of a dominant design

Regardless of the size of the technical field, the success of a technology conflict is influenced by two major groupings of elements: firm-level factors and environmental factors.

1.6.2 Firm-level factors

- **Technological advantage:** As the name suggests, this term is about the technological superiority of the product that the firm offers. Other factors being similar, technologically superior firms will most probably become the dominant ones. Thus, past experiences show that technological superiority does not always mean that the firm will be the dominant, other factors are also relevant.

- **Complementary assets and reliability:** Teece (1986) mentions the importance of complementary assets in the growth of firms. Complementary assets are assets, infrastructure, or competencies that are required to enable the successful commercialization and marketing of a technological invention but are not intrinsically related to that innovation. Credibility is yet another important aspect, as past achievements of the firms tend to make the customers believe in the new products that the company releases to the public. In the same manner, past mistakes can also be projected to the future of the firms.

- **Strategic maneuvering of the firm:** This factor describes the main elements of strategy available for the firm which is involved in the battle for dominance. Existing empirical and theoretic literature depicts four main elements: Industry entry timing, pricing strategy, licensing policy or how the firm manages its relationship with complementary services and goods, and finally, intensity and type of marketing the firm follows to present the product to the public.

1.6.3 Environmental factors

- **Governmental Regulations and institutional interventions:** The government in some cases intervenes with the use of certain technology and applies regulations for certain reasons. A recent example that applies to Europe is making GSM technology in telecommunication a standard despite among competing technologies CDMA was technologically superior. Government not only intervenes by regulations but also in some cases purchase of a product by the government in the early stages contributes to making this product the dominant one in the market. Additionally, private institutions such as industry associations such as American National Standards Institute (ANSI) can influence the technology which will become dominant.

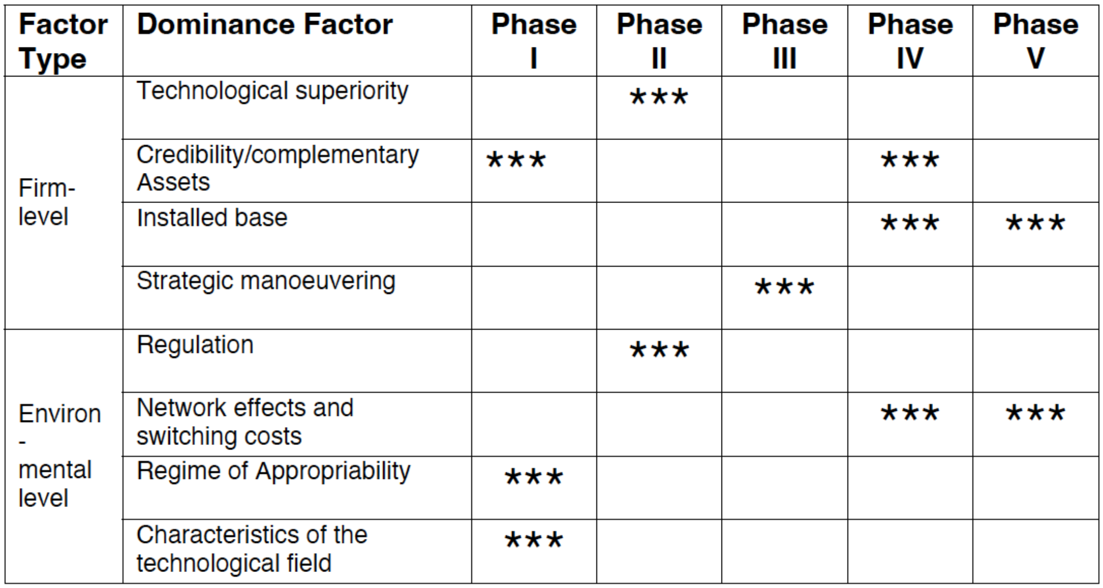
- **Size of the base of a firm:** Although the installed base of a firm depends on the outcome of the firm’s relative positioning in the other variables, it can have a significant effect on the demand of customers if the network effects are present in the field (Katz, Shapiro, 1985). The size of the installed base is associated with higher rates of adoption of the firm’s product.

- **Network effects and switching costs:** In companies’ contexts, network effects develop as a result of consumption complementarities, in which the value derived by a consumer is modified by the total number of consumers enrolled in the same network. In other words, as the network's user base grows, the demand curve swings upward. The literature explains two types of network effects which are direct and indirect (Katz and Shapiro, 1985). Direct network effects occur simply because when a client joins a network, a new network connection is established for all existing customers. Indirect network effects emerge from increased demand for supplementary items or services, such as specialist training, after-sales assistance, and compatible software. Switching costs can also impact a business's ability to recruit consumers and grow or keep its installed base. Switching costs may develop as a result of network effects or independently of them. For example, most observers agree that network effects are weak for end-users of wireless technologies: users of any network may communicate effortlessly with users of other networks since network operators have made all networks interoperable.

- **Appropriability regime:** The regime of appropriability defines how a company protects its intellectual property and knowledge from imitators. It has been determined that the factors of the business environment, excluding firm and market structure, control the ability of a company to capture the rents associated with innovation under the regime of appropriability (Teece, 1986). It is a vital aspect for the companies that compete for dominance.

- **Characteristics specific to the field:** This environmental factor defines how the market is structured and how this technological field moves on. Technological domains are filled by communities of researchers in specific disciplines as well as businesses that operate across the entire value system into which the new product is to be integrated. It has been demonstrated that research communities respond to unique dynamics, rules of engagement, and practices of information exchange that transcend the institutions in which the researchers operate (Garud, Rappa, 1995). A good example of this can be open standards that are present in the community of software developers. Along with the features of the research community, the value system structure of the industry can influence the ability of the many sponsor firms to advocate for their technology alternatives.

In conclusion, F. Suarez's framework Fig 4. combines dominance factors with the phases of development of the technology in a manner that usually happens at the battles for dominance. The framework demonstrates the key factors that affect the dominance of technology in each stage of its development.



Factor Type	Dominance Factor	Phase I	Phase II	Phase III	Phase IV	Phase V
Firm-level	Technological superiority		***			
	Credibility/complementary Assets	***			***	
	Installed base				***	***
	Strategic manoeuvering			***		
Environ - mental level	Regulation		***			
	Network effects and switching costs				***	***
	Regime of Appropriability	***				
	Characteristics of the technological field	***				

Fig 4. Key factors of success at every stage of the dominance process (F. Suarez)

2. Data Centers from the technological point of view

2.1 Data Centers

An organization's common IT operations and equipment are centralized in a data center, which stores, processes, and disseminates data and applications. Data centers are crucial to daily operations since they store an organization's most critical and proprietary assets. Since data centers are critical to any company's operations, security and dependability must be a key priority. Before the public cloud, data centers were tightly managed physical infrastructures. Most modern data center infrastructures have developed from on-premises physical servers to virtualized infrastructure that supports applications and workloads across multi-cloud environments, except where regulatory restrictions demand an on-premises data center without internet access. While data centers are frequently referred to as a single entity, they are constituted of a variety of technical components. These can be classified into three types:

- **Computer:** Memory and processing power required to operate applications, which is often provided by high-end servers.
- **Storage**: Enterprise-critical data is often stored in a data center on a variety of media, from tape to solid-state drives, with several backups.
- **Networking**: Interconnections between data center components and the outside world, such as routers, switches, and application delivery controllers.

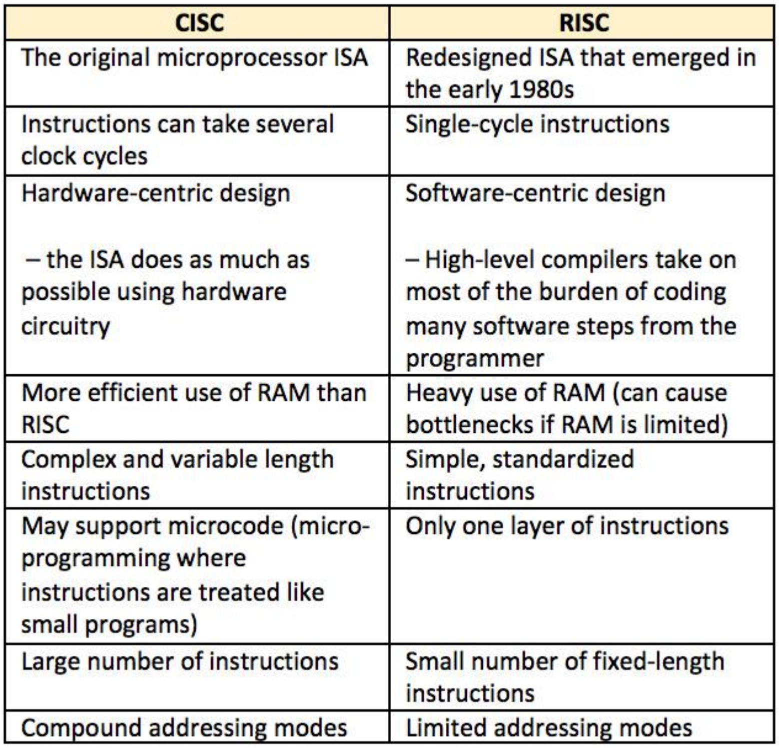
These components are vital to store and manage the resources that are important for the companies to run their continuous operations. For this reason, reliability, efficiency, and security are the main considerations. Additionally, to feed and cool down these energy-hungry facilities' advanced power systems, uninterruptable power supplies (UPS) are needed. This thesis focuses on computing components of the data centers, and specifically the chipset architecture of the computers used in data centers. Currently, CISC x86 architecture, the dominant design in the field, is being challenged by RISC architecture which in this thesis is proposed to have the potential to become the dominant design shortly. As server computers have the same infrastructure and working principle as regular computers, except the fact that server computers' components are more specialized for the required operations, we will discuss the working principle of the computers and their components in general.

2.1. Instruction Set Architecture

In the abstract concept of a computer, an Instruction Set Architecture (ISA) specifies how the CPU is controlled by software. When it comes to outlining what a processor can do and how it does it, the ISA serves as a bridge between the hardware and the software. User interaction with the hardware is only possible through the usage of the ISA. Since the assembly language programmer, compiler author, and application programmer can see it, it's like a programming manual. The ISA specifies the data types that can be used, the registers, how the hardware handles main memory, critical features (such as virtual memory), which instructions a microprocessor can execute, and the input/output paradigm of numerous ISA implementations. Additional instructions, features, and support for larger addresses and data values can be added to the ISA.

There are two types of Computers by Instruction Set Architectures: Reduced Instruction Set Computers (RISC) and Complex Instruction Set Computers (CISC).

RISC stands for Reduced Instruction Set Computer, which is a condensed form of its forerunner, the Complex Instruction Set Computer (CISC). CISC is a word that has been used to describe an architecture that differs from the RISC design; however, it did not exist at the start of processors. Many people believe that RISC is an advance over CISC in terms of performance. Architectures might be better or worse depending on the situation, hence there isn't a perfect solution. RISC-based machines execute a single instruction every clock cycle. CISC devices allow for instructions that take more than one cycle to accomplish. To run a single instruction on a RISC architecture that would take several instructions on a CISC design, multiple instructions would be required. It will need more RAM to maintain values when each instruction is loaded and then acted upon, and then a new one is loaded again, using the RISC design. With a single instruction, the CISC architecture may do all the same memory operations. However, RISC architecture consumes more RAM but executes one instruction each clock cycle, making it perfect for pipelining. RISC focuses on cycles per instruction, whereas CISC focuses on the number of instructions in each program. The amount of time it takes to execute each clock cycle, the number of cycles it takes to execute instructions, and the number of instructions in each program is all factors in determining a processor's speed. Its emphasis on huge program code sizes is evident in the RISC architecture as multiple steps in RISC equate to one instruction in CISC. It stresses software above hardware in the reduced instruction set architecture compilers and consequently, codes should be written with a smaller number of instructions to run on the reduced instruction set. It is possible to implement a greater number of instructions and more complicated ones with a complex instruction set architectures since the hardware has more transistors. In Figure 5 are given major differences between the two architectures:



CISC	RISC
The original microprocessor ISA	Redesigned ISA that emerged in the early 1980s
Instructions can take several clock cycles	Single-cycle instructions
Hardware-centric design – the ISA does as much as possible using hardware circuitry	Software-centric design – High-level compilers take on most of the burden of coding many software steps from the programmer
More efficient use of RAM than RISC	Heavy use of RAM (can cause bottlenecks if RAM is limited)
Complex and variable length instructions	Simple, standardized instructions
May support microcode (micro-programming where instructions are treated like small programs)	Only one layer of instructions
Large number of instructions	Small number of fixed-length instructions
Compound addressing modes	Limited addressing modes

Fig 5. Table of differences between CISC and RISC ISAs.

2.1. History of Computers with CISC architecture

All computers and digital devices use binary language – the digits 0 and 1 for operating and storing data. This paradigm comes from the architecture of the computers, which consists of thousands of transistors that can be turned “on” and “off”, thus, having only two states. Computers and other electronic systems work faster and more efficiently using the binary system because the system's use of only two numbers is easy to duplicate with an on/off system. Every letter, number, and symbol is represented as an 8-bit binary number. For example, the capital letter “B” is 01000010. The first prototype of a binary system in electromechanical systems was created by George Stibitz in November 1937. He built a binary adder out of light bulbs, batteries, relays, and metal strips cut from tin cans. This device was similar to a theoretical design described a few months earlier by Claude Shannon in his master's thesis.

Stibitz's "Model K" was the first electromechanical computer built. In 1939 Stibitz and Samuel Williams from Bell Labs in New York City began construction of Bell Labs Model I which was called “the first electromechanical computer for routine use.” It used telephone relays and coded decimal numbers as groups of four binary digits (bits) each. After years of developments in the field and mainly by the implementation of transistors, personal Computers being one of the main tools of a modern human had started to be truly industrialized and commercialized since 1977, with the introduction of mass-produced personal computers developed by Apple Computer Inc (now Apple inc.) with its Apple II which produced brilliant colors graphics for the time when connected to a color TV, Tandy Radio Shack with its TRS-80 that had Z80 microprocessor, video display, 4 KB of memory, a built-in BASIC programming language interpreter, cassette storage, and easy-to-understand manuals that assumed no prior knowledge on the part of the user. The TRS-80 proved popular with schools, as well as for home use. The TRS-80 line of computers later included color, portable, and handheld versions, and Commodore Business Machines with its PET computer that included either 4 or 8 KB of memory, a built-in cassette tape drive, and a membrane keyboard. The PET was popular with schools and for use as a home computer. It used a MOS Technologies 6502 microprocessor running at 1 MHz. These computers used 8-bit microprocessors that process information in groups of eight binary digits at a time which made them small and reasonably priced to be acquired by individuals for daily use in their houses, small and medium-sized businesses, primary and secondary schools. After some years, IBM Corporation came out with IBM PC which was the fastest machine among its rivals. IBM PC was using Intel 8088 CPU and its huge competitiveness was due to the invention of state-of-the-art microprocessors by Intel making Complex Instruction Set Computer (CISC) architecture a dominant design. IBM PC became the best-selling personal computer in the market at such a level that other personal computers using Intel microchips and MS-DOS systems became known as “IBM Compatibles”.

2.2. History of RISC architecture

In the late 1970s, IBM researcher John Cocke and his colleagues created the prototype computer to employ the RISC architecture. Cocke was awarded the Turing Award in 1987, the US National Medal of Science in 1994, and the US National Medal of Technology in 1991 for his contributions to computer science and technology. Cocke and his team reduced the size of the instruction set, eliminating certain instructions that were seldom used. "We knew we wanted a computer with a simple architecture and a set of simple instructions that could be executed in a single machine cycle—making the resulting machine significantly more efficient than possible with other, more complex computer designs," Cocke 1987.

The CPU could only execute a restricted set of instructions with the new design, but it could do so considerably faster because the instructions were so simple. It was possible to finish every task with only a single machine cycle (or electrical pulse), but with CISC, many tasks required multiple machine cycles and hence took at least twice as long to complete. Pipelining was made possible since each command was executed in the same amount of time. An assembly line-style method of running numerous instructions at once could be achieved through the use of pipelining. For instance, one instruction may be obtained, another may be decoded, a third may be performed, and a fourth may be used to write the result. The overall workload's throughput was improved thanks to the parallel processing of each stage. In addition, only load and store instructions could access external memory; all other operations were restricted to using internal registers. Faster computations were possible because of the new processor's simpler design.

2.3. Servers

Using a network, a server may make data, services, and applications accessible to other computers, which are referred to as "clients." Servers are computers that act as intermediaries between clients and the resources they need. Any desktop computer may be used as a Server since it has the necessary hardware and software to perform this function. Web, mail, and virtual servers are only a few of the many subcategories. The same resources can be used by two different systems at the same time. Devices are now capable of serving and receiving data in tandem. Mainframes and minicomputers were some of the first servers in use. As the name indicates, minicomputers were smaller than mainframe computers. Microcomputers are now mostly worthless because they've grown far larger than desktop computers as technology has progressed over the years. When these servers were first introduced, they were connected to clients known as terminals that did not do any calculations which were referred to as dumb terminals. They were meant to accept input from a keyboard or card reader, and then send the results of any computations to a display screen or printer. The computations were carried out on the server. It was not uncommon for a single powerful computer to be connected to a group of less powerful client computers through a network in the latter days of the Internet. The client-server paradigm is a common network design in which both the client computer and the server are capable of computing, but some responsibilities are outsourced to the server. Mainframe was a part of earlier computer designs, such as the mainframe-terminal paradigm, even if the phrase was not used. The definition of a server has evolved along with technological advancements. In today's world, a server may be anything from a piece of software operating on a computer to an entire network of computers. Virtual servers are the most common term for these servers. To begin with, virtual servers were employed to increase the number of server operations that a single physical server could do. It is a common practice nowadays to have a third-party run a virtual server on hardware linked to the Internet, known as cloud computing. For example, a mail server accepts and saves incoming emails before sending them to the appropriate client. As a file and print server, a server keeps files and accepts print jobs from clients and delivers them to a network-attached printer, among other duties.

2.4. Hardware Components of a Server computer

The majority of server computers in data centers require almost the same components to operate depending on the type and the purpose of the machine. It’s worth mentioning that the overall architectures of personal computers and server computers are almost the same, the difference however lies in the individual components and software solutions. The main components of a computer are:

- **Processor:** It executes instructions from software and hardware.
- **Memory:** It is the primary memory for data transfer between the CPU and storage.
- **Motherboard:** It is the part that connects all other parts or components of a computer.
- **Storage Device:** It permanently stores the data, e.g., hard drive.
- **Input Device:** It allows you to communicate with the computer or to input data, e.g., a keyboard.
- **Output Device:** It enables you to see the output, e.g., monitor.
- **Power Supply Unit (PSU)**: Delivers the required power to the computer and in the case of laptops PSU is substituted by a battery and charging the unit with dedicated power delivery.

2.3.1. Central Processing Unit (CPU)

A CPU is also known as a processor, central processor, or microprocessor. Being an essential part of a computer, it performs all of its vital activities. It receives instructions from both the hardware and the active software and responds by generating output in the according way. It contains all of the most critical software, including operating systems and applications. In addition, the CPU aids in the communication between input and output devices. The CPU is sometimes referred to be the computer's brain because of these characteristics.

CPU is installed or inserted into a CPU socket located on the motherboard. The CPU is kept cool by a heat sink that removes heat from the system and prevents overheating. Generally, a CPU has three components:

- ALU (Arithmetic Logic Unit)
- Control Unit
- Memory or Storage Unit

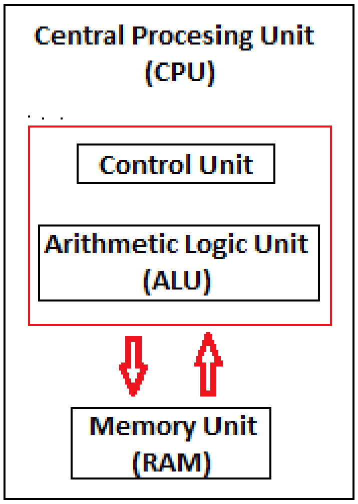


Figure 2.1 Components of a CPU

Control Unit: Electrical signals guide the computer system for the execution of previously-stored instructions in the control unit's circuitry. To begin, it reads instructions from memory, which it decodes and then executes. As a result, it controls and directs all parts of the computer. The primary responsibility of the Control Unit is to maintain and manage data flow throughout the processor. It is not involved in data processing or storage.

ALU: Math and logic operations are performed by this logic unit, which we refer to as an ALU. Addition, subtraction, multiplication, division, and comparison are all arithmetic functions. Data selection, comparison, and fusion are all examples of logical functions. A CPU can have many ALUs. ALUs can also be utilized to keep track of timers that aid in the operation of the computer.

Memory or Storage Unit/ Registers: It is called Random access memory (RAM). In addition to data, programs, and intermediate and outcomes of processing, RAM serves as a temporary storage area for these items. As a result, it serves as a temporary data storage area that is used to run the computer.

CPU Clock speed: It is the number of instructions that a CPU or a processor can process in a second that is known as its "clock speed." Gigahertz is the unit of measurement for this type of frequency. If a CPU has an effective clock speed of 3.5 gigahertz, it can execute 3.5 billion instructions per second.

2.3.2. Computer Memory

Input and instructions are stored in computer memory, which processes raw data and generates output. Cells make up the majority of the computer's storage space. In a microprocessor, each cell has an individual address that ranges from 0 to the capacity of the memory, minus one.

Computer memory is of two types: Volatile (RAM) and Non-volatile (ROM). A hard disk is considered storage, rather than a form of memory.

If we divide memory into categories on behalf of space or location, it is of four types:

- Register memory
- Cache memory
- Primary memory
- Secondary memory

2.3.3. Register Memory

Register memory is the smallest and fastest memory in a computer. Unlike the main memory, it is located in the CPU in registers, which are the smallest data storage components. The CPU uses a register to store frequently used data, commands, and memory location. CPU instructions can be found in these files. Before it can be processed, all data must travel through registers. CPUs use them to process the data that users enter.

32 to 64 bits of data can be stored in a register. The number and size (in bits) of internal registers determine a CPU's speed. Depending on their intended purpose, registers can take on a variety of distinct kinds. Accumulator or AC, Data Register or DR, the Address Register or AR, Program Counter (PC), I/O Address Register, and more are some of the most commonly utilized Registers.

2.3.4. Cache Memory

It is a type of memory that is both smaller and faster than the main memory (RAM). In comparison to the primary memory, it may be accessed more quickly by the CPU. As a result, it serves as a synchronizer and performance enhancer for high-speed CPUs.

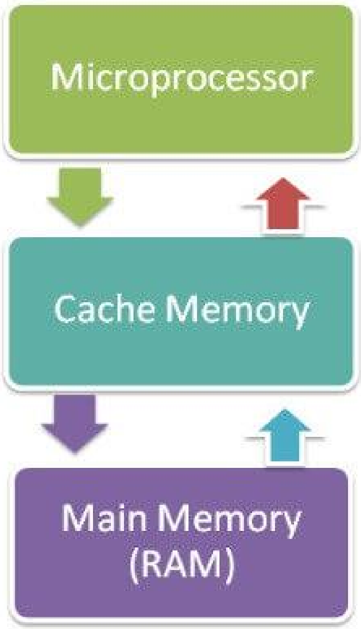


Figure 2.2 The role of cache memory

Only the CPU has access to the cache memory. It could be a section of main memory set aside for it or a storage device external to the CPU. It contains the most frequently utilized data and programs. As a result, the data is always readily available to the CPU when it requires it. For example, the CPU doesn't need to use primary memory if it can retrieve the necessary data or instructions in its cache (RAM). As a result, it improves system performance by acting as a buffer between RAM and the CPU.

2.3.5. Primary Memory

There are two types of primary memory: RAM (Volatile Memory) and ROM (Non-Volatile Memory).

RAM (Volatile Memory)

RAM, which stands for Random Access Memory, is a physical device that is often found on a computer's motherboard and serves as the CPU's internal memory. When you turn on the computer, it permits the CPU to store data, programs, and program results. It is a computer's read-only and read-write memory, so data can be stored in it and retrieved from it. In other words, RAM is not a long-term storage device for data or instructions, but rather, a temporary one. For example, if you reboot your computer, the data and instructions from the hard disk are saved in RAM; if you start a software the operating system (OS) and program are loaded into RAM from an HDD or SSD. The CPU makes use of this information to complete the tasks at hand. When you turn off the computer, the data in the RAM is lost forever. So, the data is retained in the RAM as long as the computer is running and is erased when the computer is shut down. The advantage of storing data in RAM is that reading data from RAM is much faster than reading data from a hard disk. RAM is analogous to a person's short-term memory, whereas hard disk storage is analogous to a person's long-term memory. Short-term memory recalls things for a short period, and long-term memory remembers for a long period. Information stored in the brain's long-term memory can be used to refresh short-term memory. A computer works in the same way; when the RAM is full, the CPU goes to the hard disk to overwrite the old data in RAM with fresh data. It's similar to reusable scratch paper on which you can scribble notes, figures, and other information using a pencil. When you run out of space on paper, you can erase what you no longer need; RAM works in the same way; when it fills up, the superfluous data on the RAM is deleted, and it is replaced with new data from the hard disk that is required for the current processes. RAM can be in the form of a single chip put on the motherboard or numerous chips mounted on a tiny board attached to the motherboard. It is a computer's primary memory. When compared to other types of memory, such as a hard disk drive (HDD), solid-state drive (SSD), an optical drive, it is faster to write to and read from. The size or storage capacity of a computer's RAM has the greatest impact on its performance. It will perform slower if it does not have enough RAM (random access memory) to run the operating system and software packages. As a result, the more RAM a computer has, the faster it will operate. RAM information is accessed at random, rather than sequentially, as it is on a CD or hard disk. As a result, its access time is substantially faster.

Types of RAM:

Integrated RAM chips can be of two types:

1. Static RAM (SRAM):
2. Dynamic RAM (DRAM):

Both types of RAM are volatile, as both lose their content when the power is turned off.

1) Static RAM:

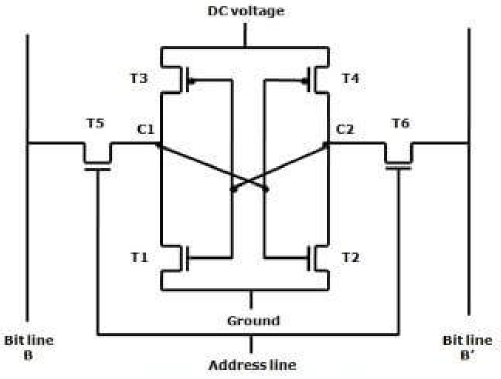


Figure 2.3 Static Ram Cell

Static RAM (SRAM) is a sort of random access memory that preserves its state for data bits or holds data for as long as power is applied to it. It is made up of memory cells and is known as static RAM since it does not need to be refreshed regularly, unlike dynamic RAM, because it does not require electricity to avoid leaking. As a result, it is faster than DRAM. It has a unique arrangement of transistors that results in a flip-flop, which is a form of the memory cell. One piece of data is stored in one memory cell. The majority of current SRAM memory cells are composed of six CMOS transistors but lack capacitors. SRAM chip access times can be as short as 10 nanoseconds. In contrast, the access time in DRAM is typically greater than 50 nanoseconds. Furthermore, because it does not wait between accesses, its cycle time is substantially shorter than that of DRAM. Because of the benefits of using SRAM, it is generally utilized for system cache memory, high-speed registers, and tiny memory banks such as a frame buffer on graphics cards. The Static RAM is quick because its circuit's six transistor arrangement keeps the flow of current in one way or the other (0 or 1). Without waiting for the capacitor to fill or drain, the 0 or 1 state can be written and read quickly. Unlike early asynchronous static RAM chips, which performed read and write operations sequentially, current synchronous static RAM chips overlap read and write operations. The disadvantage of static RAM is that its memory cells take up more space on a chip than DRAM memory cells for the same amount of storage capacity (memory) since it has more pieces than DRAM. As a result, it has less memory per chip

2) Dynamic RAM:

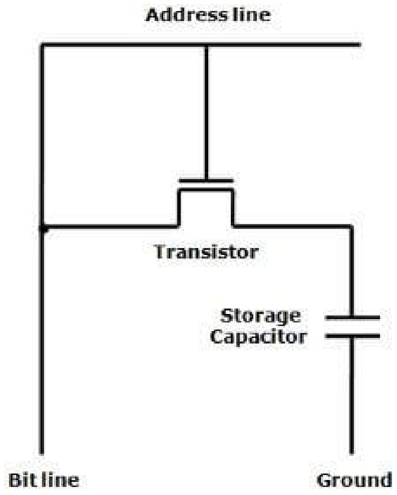


Figure 2.4 Dynamic Ram Cell

Memory cells are also used in dynamic RAM (DRAM). It is an integrated circuit (IC) composed of millions of incredibly small transistors and capacitors, and each transistor is paired up with a capacitor to form a highly compact memory cell, allowing millions of them to fit on a single memory chip. As a result, a DRAM memory cell has one transistor and one capacitor, and each cell represents or stores a single bit of data in its capacitor within an integrated circuit.

This bit of information or data is stored in the capacitor as a 0 or a 1. The transistor, which is also present in the cell, functions as a switch, allowing the memory chip's electric circuit to read the capacitor and change its state. To preserve the charge in the capacitor, it must be recharged at regular intervals. This is why it is termed dynamic RAM; it must be updated regularly to keep its data or it will forget what it is holding. This is accomplished by connecting the memory to a refresh circuit, which rewrites the data hundreds of times per second. DRAM has an access time of about 60 nanoseconds. A capacitor can be thought of as a box that stores electrons. The box is filled with electrons to store a "?1?" in the memory cell. To store a "?0?", however, it is emptied. The box has a leak, which is a disadvantage. The entire box becomes empty in a matter of milliseconds. So, for dynamic memory to function, the CPU or memory controller must replenish all capacitors before they discharge. To accomplish this, the memory controller reads memory and then writes it back. This is referred to as refreshing the memory, and it occurs automatically thousands of times every second. As a result, this sort of RAM must be constantly replenished at all times.

Types of DRAM:

i) Asynchronous DRAM:

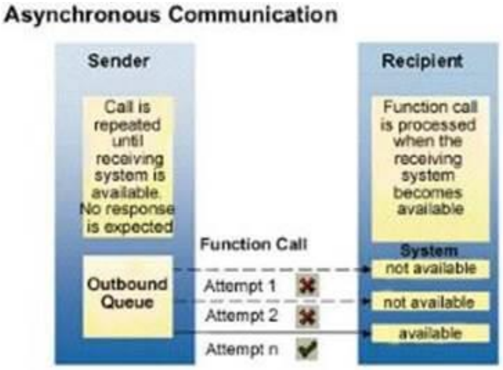

Asynchronous Communication
Sender
Call is repeated until receiving system is available. No response is expected
Outbound Queue
Function Call
Attempt 1
Attempt 2
Attempt n
Recipient
Function call is processed when the receiving system becomes available
System
not available
not available
available

Figure 2.5 Asynchronous DRAM

This type of DRAM is not in sync with the CPU clock. As a result, the disadvantage of this sort of RAM is that the CPU cannot predict the exact timing at which data from the RAM will be available on the input-output bus. The following version of RAM, known as synchronous DRAM, overcame this issue.

ii) Synchronous DRAM:

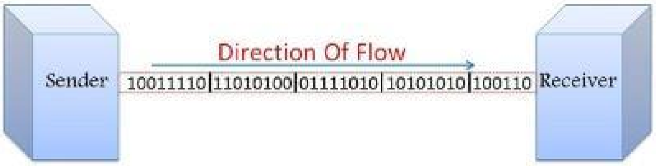


Figure 2.6 Synchronous DRAM

SDRAM (Synchronous DRAM) debuted in late 1996. The RAM in SDRAM was synced with the CPU clock. It enabled the CPU, or more precisely, the memory controller, to determine the exact clock cycle or timing, or the number of cycles after which data will be available on the bus. As a result, the CPU is not required for memory access, so memory read and write speeds can be enhanced. Because data is sent only at each rising edge of the clock cycle, SDRAM is sometimes known as a single data rate SDRAM (SDR SDRAM).

iii) DDR SDRAM:

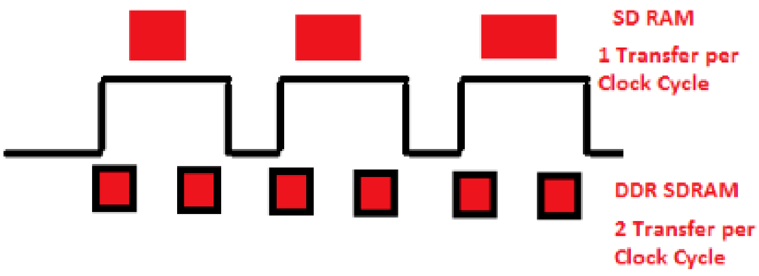


Figure 2.7

The DDR RAM is the next generation of a synchronous DRAM. It was created to address the constraints of SDRAM and was first used in PC memory in the year 2000. Data is transferred twice during each clock cycle in DDR SDRAM (DDR RAM), once during the positive edge (rising edge) and once during the negative edge (falling edge). As a result, it is referred to as double data rate SDRAM. DDR SDRAM comes in several generations, including DDR1, DDR2, DDR3, and DDR4. Nowadays, the memory that we use inside our desktops, laptops, mobile devices, and so on is usually DDR3 or DDR4 RAM. Types of DDR SDRAM are the following:

a) DDR1 SDRAM:

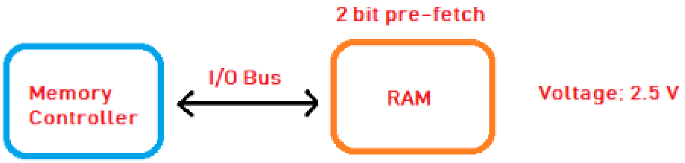


Figure 2.8 DDR1 SDRAM

DDR1 SDRAM was the first advanced SDRAM version. The voltage in this RAM was decreased from 3.3 V to 2.5 V. The data is sent on both the rising and falling edges of the clock cycle. As a result, instead of one bit being pre-fetched in each clock cycle, two bits are pre-fetched, which is known as the two-bit pre-fetch. It is typically used in the frequency range of 133 to 200 MHz. Furthermore, because data is sent during both the rising and falling edges, the data rate at the input-output bus is double the clock frequency. So, if a DDR1 RAM operates at 133 MHz, the data rate would be doubled, resulting in a 266 Mega transfer per second data rate.

ii) DDRII SDRAM:

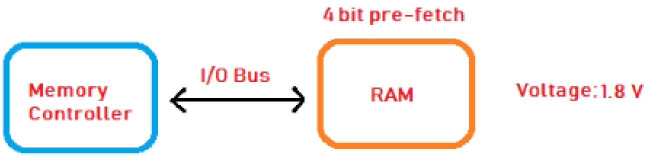


Figure 2.9 DDRII SDRAM

DDR2 is a more improved variant of DDR1. It runs on 1.8 volts rather than 2.5 volts. Its data rate is double that of the previous generation due to an increase in the number of bits pre-fetched during each cycle; 4 bits instead of 2 bits are pre-fetched. This RAM's internal bus width has been doubled. For example, if the input-output bus is 64 bits wide, the internal bus will be 128 bits wide. As a result, a single cycle may handle twice as much data.

iii) DDR3 SDRAM:

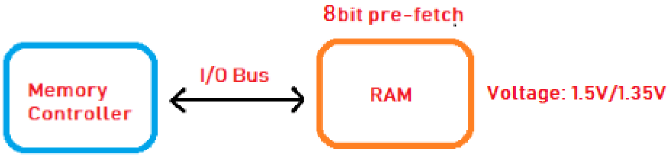


Figure 2.10 DDRIII SDRAM

In this version, the voltage is dropped even further, from 1.8 V to 1.5 V, which is a significant reduction. Because the number of bits that are pre-fetched has been increased from 4 bits to 8 bits, the data rate of the new generation RAM is double that of the previous generation RAM. It is possible to say that the internal data bus width of RAM has been raised by two times compared to the previous generation.

iv) DDR4 SDRAM:

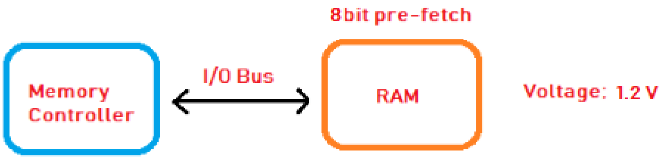


Figure 2.11 DDRIV SDRAM

In this version, the operating voltage is further reduced from 1.5 V to 1.2 V, but the number of bits that can be pre-fetched is the same as the previous generation; 8 bits per cycle. The Internal clock frequency of the RAM is double of the previous version. If you are operating at 400 MHz the clock frequency of the input-output bus would be four times, 1600 MHz and the transfer rate would be equal to 3200 Mega transfers per second.

ROM (Non-volatile Memory)

Any storage medium that can only store data in the read-only mode is known as a read-only memory, or ROM for short. Along with random access memory, it acts as the computer's primary memory unit. It is called "read-only memory" the data and programs stored on it are only readable. Thus, It can only read words that have been permanently saved on the computer.

In the manufacturing process of a ROM, the manufacturer installs the programs. In the future, the ROM's content cannot be modified, rewritten, or removed. However, It is possible to modify the information in several types of ROMs. There is an electrical fuse in the read-only memory (ROM) capable of operating in a certain wiring arrangement. During the manufacturing process, the designer defines the binary data stored on the chip, which is then placed into the unit to achieve the necessary connection pattern. Even if the power is turned off, it will stay in the device. Because it holds data even after the power is switched off or the computer is shut down, it is a non-volatile memory. Programming the ROM is the act of adding bits of information to the RAM, where the bits are stored in the hardware configuration of the device. . Simple ROM is the cart used in video game consoles to run several non-modifiable games on the device. Personal computers and other electronic gadgets, such as telephones, tablets, televisions, and air conditioners also contain permanently stored data. When booting a computer there are starting instructions stored in the ROM that are needed to start the computer, therefore it takes a while to show up. The computer's operating system begins with the computer's boot process. The operating system is loaded into the primary memory of the machine during this process. The computer's CPU starts the booting process using the BIOS program, which is also stored in the computer's read-only memory. BIOS connects the computer with its operating system. A piece of hardware, such as a keyboard, hard drive, or video card, can hold software programs called Firmware in the ROM of a computer. An electronic device's Flash ROM contains it. It teaches the gadget how to talk to and interact with other gadgets in the digital world. There are 5 types of reading only memory:

1. **Masked Read-Only Memory (MROM):**

   It is the most primitive form of read-only memory (ROM). It has become obsolete and is therefore not utilized in today's world. It is a type of physical memory device in which the manufacturer stores programs and instructions during the manufacturing process. As a result, it is coded at manufacture and cannot be edited, reprogrammed, or erased later.

2. **Programmable Read-Only Memory (PROM):**

   This type of ROM is manufactured as an empty memory to be programmed later. A specialized instrument is used to burn data inside of it once and this data is then kept forever. Because the data cannot be updated after it is programmed, it is sometimes referred to as a one-time programmable device. It is utilized in a variety of applications, including cell phones, video game consoles, medical devices, and RFID tags.

3. **Erasable and Programmable Read-Only Memory (EPROM)**

EPROMs are a type of read-only memory that can be written and erased several times. However, it uses a different technology than traditional rewritable memory architecture. Data is erased by delivering precise frequency ultraviolet light from a quartz window. To rewrite a data inside specialized program called PROM burner is used.

4. **Electrically Erasable and Programmable Read-Only Memory (EEPROM)**

EEPROM is a type of read-only memory that can be erased and reprogrammed up to 10000 times. It is also referred to as Flash EEPROM as it resembles flash memory because of its characteristics. However, instead of writing and erasing data in blocks such as in the case of flash memory, an EEPROM does it one byte at a time. This memory is used to carry the BIOS of machines.

5. **Flash ROM**

Flash ROM is a better version of EEPROM because of its data write/erase architecture that allows blocks of data to be managed which makes it more versatile. It also can be reprogrammed without removing it from the computer. It is also reasonably durable for a wide temperature range and high pressure. This type of memory is used in USB flash drives, Mp3 players, digital cameras, and solid-state drives (SSDs).

2.3.6. Secondary Memory

The secondary storage devices that are integrated into or linked to the computer are referred to as the computer's secondary memory. Additionally, it is referred to as external memory or supplementary storage. Through input/output actions, the secondary memory is accessible indirectly. It is non-volatile, which means that the data is retained indefinitely even when the computer is shut off or until it is rewritten or deleted. The CPU cannot access secondary memory directly. Secondary memory data must first be copied to primary memory before the CPU may access it. Secondary memory units include hard disks, solid-state drives, pen drives, and SD cards, among other storage devices.

2.3.7 Graphics Processing Unit

Graphics Processing Unit isn’t usually referred to as one of the main components of the PC architecture, however, for some operations servers require a GPU. CPU and GPU have a similar structure and working principle, however, there is a key difference that distinguishes these two. GPU is designed for parallel computing and does it much faster than a CPU does. To be clearer, CPUs jump through several tasks requiring lots of interactivities whereas, GPU solves tasks in a slower manner but in big chunks which makes it very fast for computer graphics, video games, and parallel computing tasks. Architecturally, the CPU contains just a few cores with lots of cache memory that can handle a few threads at a time. However, a GPU is made of hundreds of cores that can handle thousands of threads simultaneously.

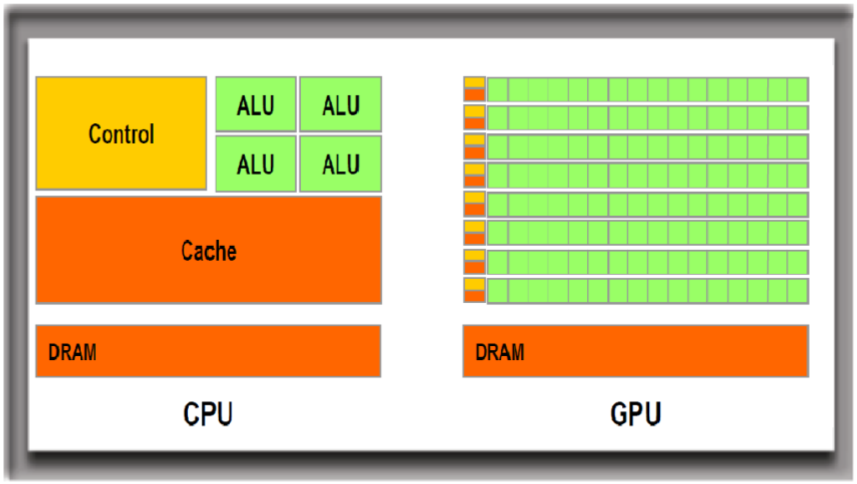


Fig 2.15 CPU versus GPU architecture

There are two types of GPUs: integrated and discrete. Integrated GPUs come embedded alongside the CPU. A discrete GPU however, is a chip that is mounted on its circuit board and is typically attached to a PCI Express slot of the PC, or in the case of laptops, it is mounted to the motherboard separately from the CPU. Currently, Integrated GPUs of x86 CPUs are not powerful enough to coup up with the needs of most computer users which makes it necessary for PC manufacturers to include discrete GPUs alongside the system to manufacture more powerful computers to meet customer needs.

3.Data Centers Market Analysis

Before diving deep into the methodology, it is vital to understand the current situation in the global market of Data Centers and assess its importance for the world. People generate more than 2.5 million gigabytes of data each day which makes a flow of $100 billion flow of funds into the data center ecosystem by a variety of organizations. This capital contributed to the growth of the industry due to a lower cost as rather than spending on data center assets, firms allocate their investments to other important factors for growth. Therefore, during recent years parallel with improvements in the Data Center industry significant shift from on-premises data facilities to cloud data services is taking a place. This shift has made the biggest Data Centers in the world grow even larger and more efficient. With the growing popularity of the Internet of Things (IoT), this rate of data production will accelerate even further. Cushman & Wakefield conducts Data Center Global Market Comparison to rank the top Data Centers of the World measuring them using a methodology that covers every aspect that may make certain data centers better than the others.

The 2022 Global Data Center Market Comparison scores each data center across 3 criteria which contains 13 categories that cover almost every aspect to fairly compare the Data Centers of the world. These criteria are real estate and physical (development pipeline, environmental risk, land price, vacancy), ecosystem advantages (cloud availability, fiber connectivity, market size, sustainability, smart cities), and political and regulatory review (government incentives, political stability, power cost, taxes). Then, these 13 criteria were weighted by their importance as mentioned on

Fig 3.1. The research control group contains data from 30 research sources, 1,162 data centers, and 38 global markets.

High-Weight	Mid-Weight	Low-Weight
Fiber Connectivity	Incentives	Power Cost
Market Size	Taxes	Land Price
Cloud Availability	Political Stability	Environmental Risk
	Vacancy	
	Development Pipeline	
	Sustainability	
	Smart Cities	

Fig. 3.1 Weighted categories that affect Data Centers credibility

To obtain the ranking results for 2022 top ten biggest markets that use data centers were considered from 38 which are Cape Town, Moscow, Athens, Abu Dhabi, Vienna, Istanbul, Hyderabad, Bangkok, Auckland, and Mombasa.

Results of the ranking:

1. Northern Virginia
2. Silicon Valley*[1] / Singapore*
4. Atlanta* / Chicago*
6. Hong Kong
7. Phoenix
8. Sydney
9. Dallas
10. Portland* / Seattle*

Data centers being power-hungry facilities have been reported to consume around 190 terawatt-hours of energy in 2020. Particularly, 41 terawatt-hours are reported to be consumed by traditional data centers, 73 terawatt-hours by non-hyper-scale cloud, and 76.2 terawatt-hours by hyper-scale cloud computers. In the United States, data centers consume more than 2% of total power which is a huge number. All of these demonstrate

[1] * Ranking Tie

how important are data centers for us and with growing worldwide data exchange these numbers will grow too. In parallel with the increasing demand, the efficiency of the Data Centers is also increasing. For measuring the efficiency of Data Centers power usage efficiency (PUE) metric is usually used. The formula for the calculation of PUE is as follows:

$$PUE = \frac{Total\ Facility\ Energy}{IT\ Equipment\ Energy} = 1 + \frac{Non\ IT\ Facility\ Energy}{IT\ Equipment\ Energy}$$

The lower the PUE the better. It has been reported that in 2007 PUE of the largest data centers was 2.5 on average. However, in 2020 this number dropped to 1.59. Governments encourage data centers to optimize their facilities and achieve lower PUE ratios.

From the financial point of view, only in 2020 data centers worldwide have made a revenue of 91.02 billion U.S. dollars and shipment of 12.15 million units have been made 11.8 of them being x86 (CISC) architecture covering 97% of the industry which makes x86 the absolute dominance in the server market. By segment, most of the revenues for high performance capturing (HPC) were captured by server operations being $11.846 million followed by $4.772 million by storage and $4.300 million by HPC cloud spending. Although in 2020 RISC-based enterprise center processors have made a revenue of $4.1billion which is only 4.5% of the total revenues, it is predicted that in 2030 this number will increase to $82 billion.

4. Methodology

4.1. Definition of the methodology

Selecting the right considerations for analyzing and predicting the future of the dominant design is a very delicate and important task as many variables are involved in predicting what the future will bring for the market. In our case CISC being a dominant design in the market is being challenged by RISC. Dislocating an already established architecture from the market is more difficult and this matter should be analyzed more precisely. This methodology is based on an outstanding integrative framework built by F.Suarez analyzing the variables affecting the dominance in the market according to the variables on each phase of the market. The key difference is, instead of focusing on the technology on the firm level, the focus on it will be on the industry level, in an already established market considering the case in CISC versus RISC battle. Assuming the previously mentioned conditions, an additional framework based on F.Suarez's work will be developed to analyze the industry from the cradle to find out if the technology has the right market requirements in the long run to achieve market dominance. The methodology will cover possible factors that affect the dominance in the market of database server chipsets within the limitations of the study. In this methodology, a direct comparison of the competing architectures and a rough estimate of the chance of the competing new architecture becoming dominant in the market will take place. The considerations will be as follows:

1. Technological Comparison
2. Complementary assets
3. Strategic situation of the industry
4. Possible Governmental Regulations and institutional interventions
5. Possible Switching costs

6. Appropriability regime
7. Customers’ point of view

In the methodology, it is assumed that technological advantage is the absolute necessity, each of the remaining factors contributes to the new technology to challenge the dominant one. To put it differently, the assumption is if the market is already established, the new technology should necessarily have a technological advantage over the established ones to have a chance to win the battle. Factors other than technological advantage will increase the likelihood of the technology becoming dominant in the market. Following the methodology bodies analyzing the industry should have a road map for understanding what the chances of success are and which factors are more important and need more resources implemented on them to achieve better results.

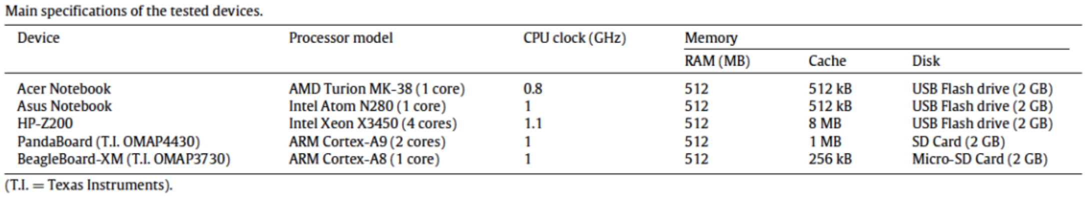
4.2. Technological Comparison

The technological advantage being the most important factor will be analyzed in more depth. In data servers even incremental improvements in efficiency multiply and results in a significant reduction in the cost because they consist of thousands of computers working altogether. To make a comparison and select the superior technology between the two direct comparisons by benchmarking will be used. To compare two architectures two different benchmarks from the literature will be analyzed.

It has long been associated with battery-powered embedded devices, but data centers and high-performance systems are now taking a closer look at power usage. Electric power consumption is a serious problem in data centers, which is one of the main concerns about them. Servers and cooling systems are the largest energy consumers, according to Mahadevan et al. Reduced power use is desirable not only for ecological reasons but also to save money on utility bills. According to the J.Hamilton model, 57% of monthly data center costs are spent on server electricity bills.

4.2.1 First Benchmark

A study that involves a direct comparison of the two architectures in the same environments has been held to measure the performances of the two systems using real-life situation tasks that are similar to what are the roles of these systems in Data Centers. These tests involve Web Server, Database Server, and Floating-point comparisons. The comparison hardware consists of two laptop PCs equipped with low-power x86 CPUs and two ARM-based development boards. The specs for each device are listed in Table 4.1



Main specifications of the tested devices.

Device	Processor model	CPU clock (GHz)	Memory		
			RAM (MB)	Cache	Disk
Acer Notebook	AMD Turion MK-38 (1 core)	0.8	512	512 kB	USB Flash drive (2 GB)
Asus Notebook	Intel Atom N280 (1 core)	1	512	512 kB	USB Flash drive (2 GB)
HP-Z200	Intel Xeon X3450 (4 cores)	1.1	512	8 MB	USB Flash drive (2 GB)
PandaBoard (T.I. OMAP4430)	ARM Cortex-A9 (2 cores)	1	512	1 MB	SD Card (2 GB)
BeagleBoard-XM (T.I. OMAP3730)	ARM Cortex-A8 (1 core)	1	512	256 kB	Micro-SD Card (2 GB)

(T.I. = Texas Instruments).

Table 4.1

PandaBoard and BeagleBoard-XM were utilized for this project as they are open-source boards. It was also necessary to compare the results to those obtained using a more typical server system. In this case, an Intel Xeon quad-core CPU-powered HP Z200 workstation PC was tested. The Linux kernels 2.6.32, 2.6.35, and 2.6.37 are installed on each machine. All systems use MySQL 5.1 for the database, while the Apache HTTP server is used as a web server. USB flash drives were used to boot all machines, while RAM was used to execute the tests to avoid any interference from disk I/O pings. The Apache HTTP web server's basic configuration was modified to support concurrent threads up to five hundred. A remote monitoring station employs remote shell commands to execute the tests and gather data. When a machine boots, its RAM is restricted to 512MB, and the CPU is designed to run at about 1 GHz. GUI is avoided and the number of executing apps is minimized. For the evaluation, a mySqlSlap was used to simulate client demand on the server. To execute the test, the webserver Apache benchmark tool was used running a direct 100 Mbit Ethernet connection between the monitoring host and the system under test. The floating-point performance was evaluated using Linpack C. Ten thousand HTTP requests or 512 SQL queries were made for each test. The monitoring device captured second-by-second data on temperature, CPU utilization, I/O latency, and power consumption throughout the experiment. For the Linpack test, the given number N is utilized to perform LU factorization benchmarks in a system of NxN matrices.

Results

- **Web Server**

Data obtained from all platforms utilized to assess the webserver is depicted in Figure 4.1. Static web pages are requested by between one and one thousand clients increasing by 25 each time during the tests. A Bezier curve approximation is used to smooth the graphs. The plots depict performances when all cores are enabled, the dashed lines show the results obtained running only on one core.

The power consumption graph shows how much power is required when a certain number of clients are active at the same time. On average, Atom uses 8.98W for two-threaded operations and 9.32W for four. As compared to the Cortex A8, Turion consumes an average of 17.38W. The Cortex A9 has a single-core power consumption of 4W and dual-core power consumption of 4.55W. When only one core is active, the Xeon consumes 64.12W, whereas when all four cores are active, it consumes 61.60W. Each device's service quality may be deducted in the response latency graph. Xeon response time is 70 milliseconds, whereas a four-core Xeon may reach a maximum response time of 145 milliseconds. The maximum delay is 549 milliseconds, with an average of 235 milliseconds, when using a single core. There are no acceptable response times for interactive web apps on the Cortex A8, which has an average of 942 milliseconds and a maximum of 1943 milliseconds when 1000 clients perform requests. A reaction time of 1224 milliseconds is the worst-case scenario for Turion, while an average response time of 577 milliseconds is the norm. Atom's worst execution time is 969 milliseconds, while its average execution time is 516 milliseconds. Arm Cortex-A9 is the second-fastest processor with two cores, with the worst value of 754 milliseconds and an average of 370 milliseconds. The worst reaction time, however, jumps to 1667 milliseconds with a single core.

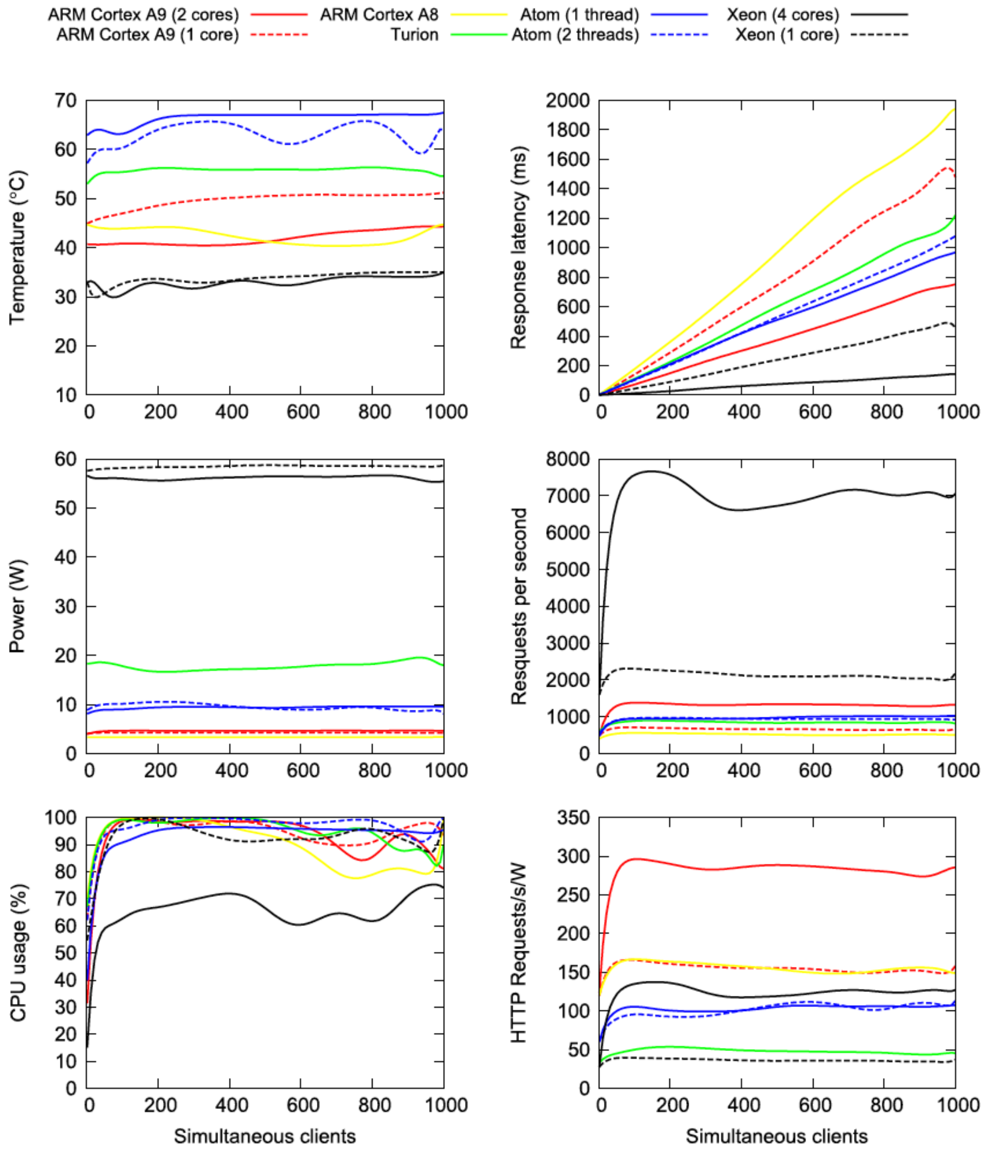


Fig 4.1

The graph HTTP Requests/s/W represents a direct divide of the requests/s by the power consumption, for each concurrent client. As expected from the investigation, this graph demonstrates that all ARM-based devices have a higher requests/s/W efficiency, even under the highest possible load. The performance per watt of a Xeon with a single core is the lowest, followed by Turion and Atom. The Cortex A8 and A9 cores of ARM have nearly identical performance per wattage when only one core is used. For example, 2 core Cortex A9 consumes 8.7 times less energy than a single-core Xeon and 2.5 times less energy than a four-core Xeon. Experimental measurements show that Cortex A9 can handle 2.5 more requests per Watt. The tests conducted for the webserver evaluation using static web pages revealed that RISC-based systems consistently outperformed CISC-based systems in terms of performance per watt.

- **Database Server**

Fig 4.2 displays charts showing the average values collected across all platforms used to evaluate the database server. Each plot in Fig 4.2 depicts the results for the MySQL server running with clients ranging from 1 to 130, with each measurement batch increasing by one client.

The power plot demonstrates that Xeon consumes more power when four cores are active than when a single core is active, as expected. The max power consumption of the dual-core Cortex-A9 is 5.34W (average of 4.52W), whereas the single-core consumes 4.63W. (average of 4.09 W). The average power consumption of an atom is 7.25 W, with a peak of 7.41 W. Turion has a maximum power output of 18.1 W and an average power output of 14.62 W. The ARM Cortex-A8 consumes an average of 2.76 W and peaks at 2.94 W. Xeon consumed a max power of 62 watts with a single core and 73.3 watts with four cores (averages of 61.5W and 67.89W respectively).

The maximum response times for database queries are as follows: Atom: 51.46 milliseconds, Turion: 17 milliseconds, Cortex-A8: 77 milliseconds, Cortex-A9 (2 cores): 59.62 milliseconds, Cortex-A9 (1 core): 46.96 milliseconds, Xeon (1 core): 9.24 milliseconds, and Xeon (4 cores): 8.57 milliseconds. The average values are as follows: Atom: 32.47 milliseconds; Turion: 6.72 milliseconds; Cortex-A8: 40.41 milliseconds; Cortex-A9 (2 cores): 27.10 milliseconds; Cortex-A9 (1 core): 25.24 milliseconds; Xeon (1 core): 4.26 milliseconds; and Xeon (4 cores): 4.16 milliseconds.

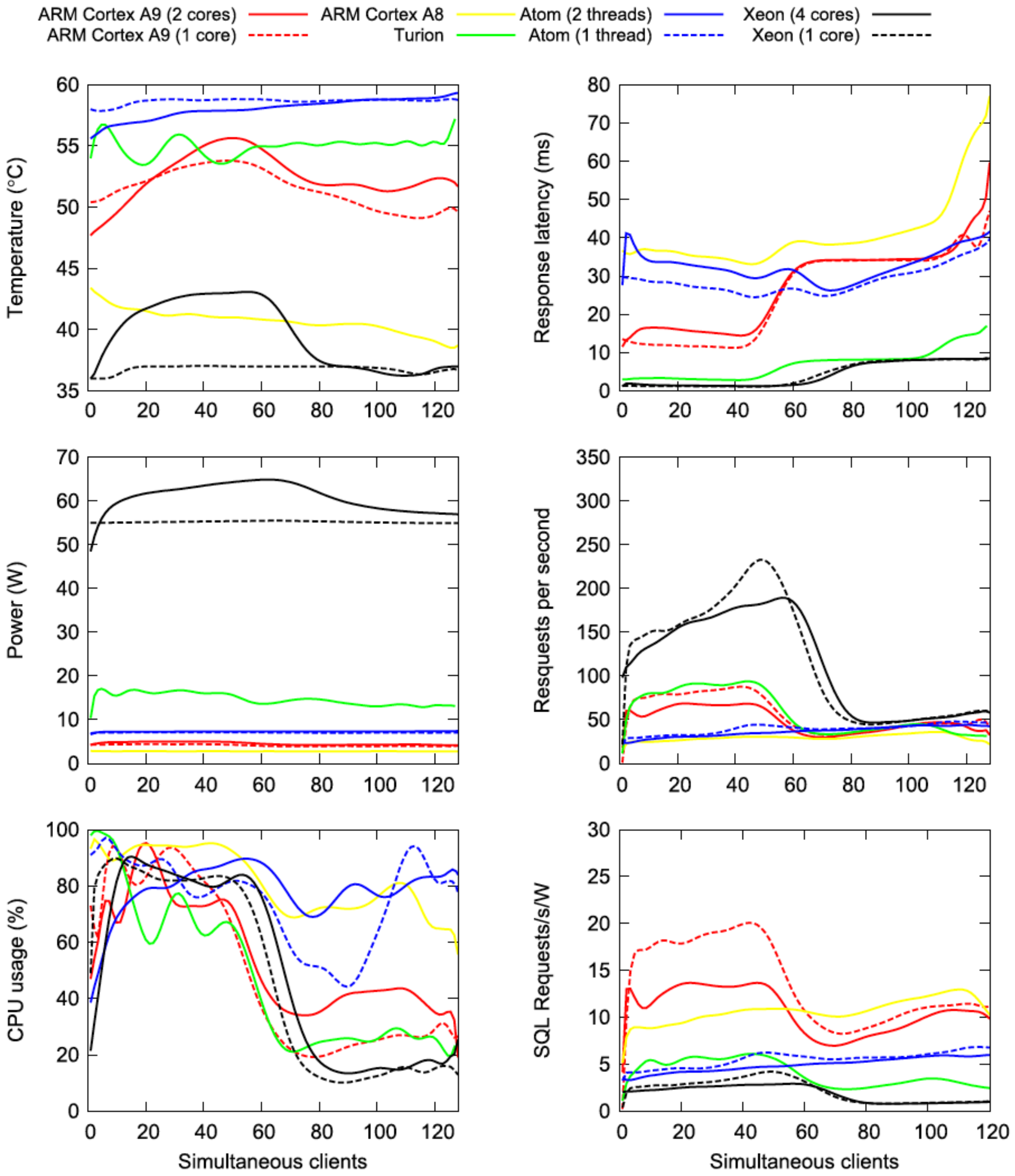


Fig 4.2

The following are the maximum sustained queries per second (rounded): Atom has 50 queries, Turion has 152, Cortex-A8 41, Cortex-A9 (2 cores) 105, Cortex-A9 (1 core) 125, Xeon (1 core) 349, and Xeon (4 cores) has 220. The average values are 36 for Atom, 57 for Turion, 29 for Cortex-A8, 49 for Cortex-A9 (2 cores), 56 for Cortex-A9 (1 core), 116 for Xeon (1 core), and 112 for Xeon (4 cores).

RISC systems outscored CISC systems in terms of performance per watt in all tests done for the database server evaluation. Additionally, their latencies were much greater than those of the Xeon CPU, as was the case with the Web Server testing. The database server faced concurrency issues, as opposed to the HTTP server, which had minimal communication within the process. This resulted in a significant improvement in performance while only one CPU core was activated. Adding more cores resulted in lower output in terms of both raw power and power consumed per watt. Concurrency and locks must be appropriately addressed to achieve improved power efficiency in multicore systems, as demonstrated by this study.

- **Floating Point**

The graphs in Figure 4.3 depict the average data collected across all platforms during the tests. Additionally, this figure contains a plot summarizing the results and displaying the MFlops per Watt rates seen in each system. Turion consumes nearly twice as much energy at 2.2 GHz as it does at 800 MHz, as the power graph demonstrates. Due to the system's overall power consumption, the increase in Xeon power consumption is rather small. The power consumption of each system is as follows: a peak of 10 W for Atom at 1.6 GHz with an average of 9.19 W, a peak of 8.22 W for Atom at 1 GHz with an average of 7.79 W, a peak of 3.58 W for ARM Cortex-A8 at 1 GHz with an average of 3.44 W, a peak of 4.54 W for ARM Cortex-A9 at 1 GHz with an average of 4.28 W, a peak of 45.1 W At 1.1 GHz, the Xeon processor consumes a maximum of 80.8 W with an average of 71.18 W and a maximum of 80.9 W with an average of 76.38 W. At 2.67 GHz, the Xeon processor consumes a maximum of 80.9 W with an average of 76.38 W. CPU utilization is always greater than 90% for all CPUs, as depicted in the graph (most of the time at 100 percent).

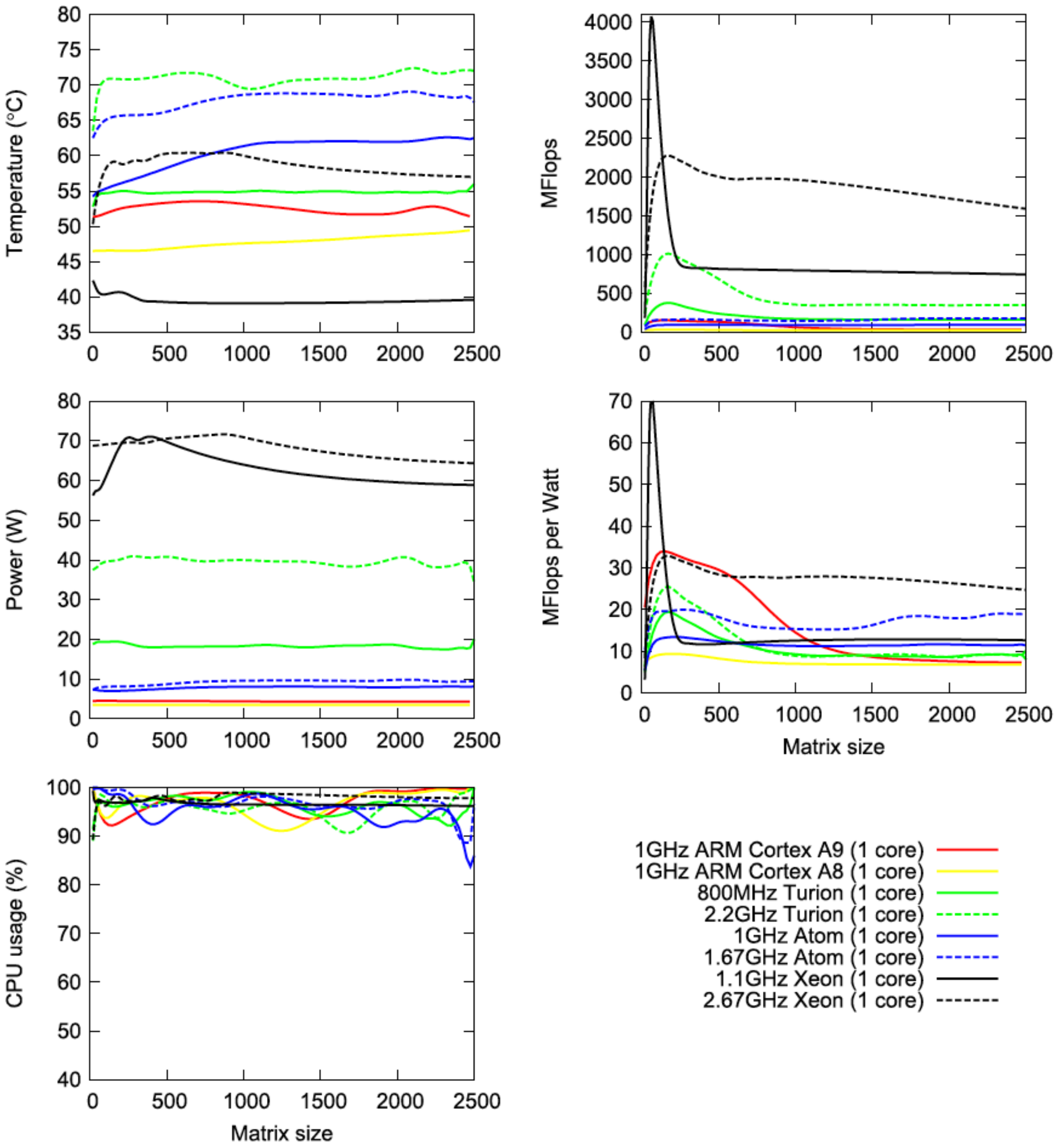


Fig 4.3

The MFlops graph displays the results of Linpack execution for each matrix size. All systems work optimally while Linpack performs LU factorization on a system of matrices with dimensions 40x40, 50x50, 60x60, 70x70, and 80x80. For these matrix sizes, Xeon provides the best performance. The following are the results: Atom (1.6 GHz) has a high of 178 MFlops and an average of 158 MFlops, while Atom (1 GHz) has a peak of 98 MFlops and an average of 91 MFlops. Cortex-A8 has a maximum performance of 33 MFlops and an average of 25 MFlops, whereas Cortex-A9 has a maximum performance of 172 MFlops and an average of 68 MFlops. At 2.2 GHz, Turion (1.1 GFlops) has a peak of 1.1 GFlops with an average of 464 MFlops while at 800 MHz, it has a peak of 450 MFlops with an average of 194 MFlops. At 1.1 GHz, Xeon achieves a peak performance of 6 GFlops with an average of 1 GFlop and a peak performance of 2.4 GFlops with an average of 1.8 GFlops at 2.67 GHz.

According to the "MFlops/Watt" graph, the Xeon at 2.67 GHz has the highest long-term power efficiency. Although the ARM Cortex-A9 has a higher power efficiency for matrices between 20x20 and 500x500, only the Xeon maintains its performance across all tested matrix sizes.

ARM-based SoCs perform well when used to build servers and clusters, even more so when measured in terms of performance per watt. According to the testing on HTTP and SQL servers, ARM devices are between three and four times more energy-efficient than x86 systems when the requests per second per Watt ratio is considered under various load conditions. The exception is the floating-point calculation, where the Cortex A9 was more efficient for a limited range of issue sizes before degrading. On the other hand, CISC processors maintained nearly constant performance in floating-point calculations compared to RISC, in terms of both performance and efficiency.

4.2.2 Second Benchmark

According to the J.Hamilton (Fig 4.6) model power directly contributes 13% of the entire cost of the data center. However, power has an indirect effect on infrastructure costs because cooling and power distribution infrastructures are designed around the maximum amount of power dissipated by servers. As a result, optimizing server energy use may save 31% of the entire data center cost. The quad-core ARM cortex A9 CPU is evaluated in this work using a Versatile Express development platform, whereas the dual-core ARM Cortex A9 processor is examined using a Tegra 200 series developer kit. The highly configurable express is comprised of a V2M-P1 motherboard and a CoreTile V2P-CA9 Express A9 MPCore logic board. The logic board features 1GB of DDR2 memory and a 400MHz Cortex A9 NEC CPU. The Tegra 200 series developer kit features a Tegra 250 CPU clocked at 1 GHz and equipped with 1GB DDR2 memory. Three benchmarks representative of common data center and server farm applications were examined on these two platforms:

- Autobench to evaluate the performance of the Apache 2.2 HTTP server
- SPECweb2005
- Erlang runs a time system.

Table 4.2 compares the performance and energy efficiency of Cortex-A9-based platforms for common server tasks to those of x86-based platforms. These are the results for Apache 2.2 serving a ten-byte static file. Although the quad-core Intel Xeon platform can handle seven times the number of requests per second as the dual-core Cortex A9, Table 4.2 demonstrates that the ARM-based processor has tenfold higher energy efficiency. The SPECweb2005 benchmark was used to assess the Tegra 250 processor's performance with more demanding web services. SPECweb2005 is a collection of three distinct workloads: support, e-commerce, and financial services. The support workload is modeled after that of a hypothetical customer assistance web service, the e-commerce workload is modeled after that of a web-based shopping system, and the banking workload is modeled after that of an online banking system. The performance of SPECweb2005 on two x86 machines and the Tegra 250 is shown in Table 4.3, while the energy efficiency of the systems is shown in Table 4.4. In this comparison, two Xeon X3360 machines are utilized as references, the second of which features an optimized disk architecture for serving the data demanded by the benchmarks. The improved Xeon X3360 platform can support approximately 33 times the number of sessions as the Tegra 250 platform but for three times the power efficiency. Finally, an Erlang-based SIP proxy is utilized as a benchmark to evaluate the researched CPUs' performance and energy efficiency on a typical telecom application present in data centers. The proxy's performance was determined by the maximum number of calls per second that the platforms could manage, and the related energy efficiency was expressed in terms of calls per joule spent. The reference x86 computer is powered by two 2.66GHz quad-core Intel Xeon L5430 processors. Table 4.5 summarizes the performance data, whereas Table 4.6 details the associated energy efficiency. The reference x86 computer handled 400 calls per second, while the quad-core Cortex A9 handled 30 calls per second with eight schedulers (SMP), as employing more schedulers than the number of physical CPUs available does not result in performance gains. This results in 25 calls per Joule energy efficiency for the quad-core Cortex A9 versus 8 calls per Joule for the Xeon system.

Machine	Request / s	Requests / J
Quad Core Intel Xeon E5430 (2.66 GHz)	33000	413
Pentium 4 (2.8GHz)	7100	80
Dual Core Cortex-A9 MPCore (1 GHz)	4600	4600
Quad Core Cortex-A9 MPCore (400 MHz)	3400	2833
Cortex-A8 (600 MHz)	760	760

Table 4.2 Ability of Apache 2.2 to serve 10 byte static files using different hardware

Machine	Ecommerce	Banking	Support
Quad Core Intel Xeon X3360 (1)	3600	2700	4200
Quad Core Intel Xeon X3360 (2)	7360	6240	7840
Dual Core Cortex-A9 MPCore (1 GHz)	230	180	220

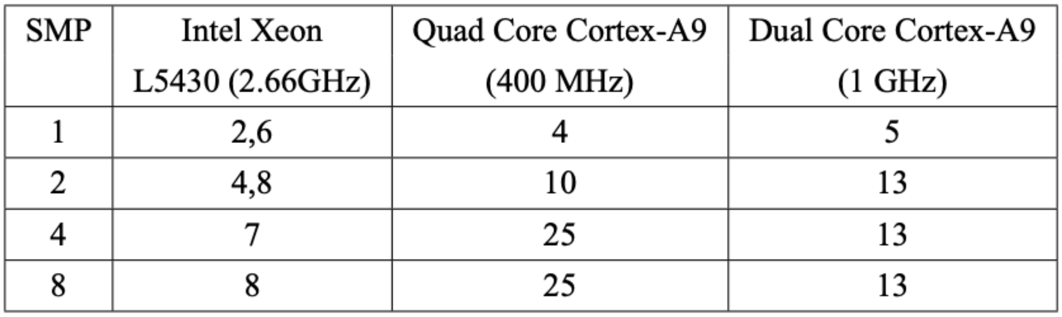
Table 4.3 Number of simultaneous sessions using different hardware

Machine	Ecommerce	Banking	Support
Quad Core Intel Xeon X3360 (1)	38	28	44
Quad Core Intel Xeon X3360 (2)	77	66	83
Dual Core Cortex-A9 MPCore	230	180	220

Table 4.4 Number of simultaneous sessions per dissipated watt

SMP	Intel Xeon L5430 (2.66GHz)	Quad Core Cortex-A9 (400 MHz)	Dual Core Cortex-A9 (1 GHz)
1	130	5	5
2	240	12	13
4	350	30	13
8	400	30	13

Table 4.5 Maximum number of calls per second handled by the Erland SIP-Proxy



SMP	Intel Xeon L5430 (2.66GHz)	Quad Core Cortex-A9 (400 MHz)	Dual Core Cortex-A9 (1 GHz)
1	2,6	4	5
2	4,8	10	13
4	7	25	13
8	8	25	13

Table 4.6 Energy efficiency in several calls per Joule

Conclusion of the Technical comparisons

Two different benchmarking results from the literature result demonstrated that the RISC system is superior and more cost-efficient than the CISC system. Although a few commercial ARM-based systems aimed at data centers and server farms have recently been introduced, significant anticipation is being placed on the forthcoming ARMv8 architectures. The industry is already developing 64-bit 3D many-core processors based on ARMv8 architectures, and the industry forecasts energy-efficient cloud data centers with several hundreds of server-in-a-chip attaining thousands of cores on a single board (S. Saponara, L. Fanucci, M. Coppola 2012). If developed to fit the Global Data Server market, it is expected that RISC will become a much more efficient, cheap, and sustainable option than CISC. So, the results give us a green tick for the Technological superiority of RISC which is as we mentioned before, the most important factor for the dominant design battle for the proposed framework.

4.3. Complementary assets

Availability of complementary assets is one of the factors highly affecting the acceptance of the new architecture, as this is what highly contributes to a design becoming dominant. As an example, Gas vehicles and the high availability of gas stations in competition with Electric vehicles for dominance can be considered which was one of the main factors that made Internal combustion engine vehicle design dominant in the market. In the case of RISC and CISC architectures, the situation in the market is different. If we observe the complementary assets of these designs, they are almost the same which makes CISC architecture reasonably vulnerable in the market. For example, memory units of the data centers remain unchanged when transferring the whole system from CISC to RISC. To put it in another way, it just takes to change the logic board and the operating software of the system to switch the CISC with RISC, and buying a RISC based system will even be cheaper than buying a CISC system as the former is an architecture suitable to provide System on a chip design Figure 4.4. To clarify, a combination of CPU, RAM, GPU can be achieved with these systems as RISC architecture makes it available, whereas, in CISC systems it was impossible to integrate RAM, GPU in the CPU. Even if there are some x86 CPUs with integrated GPU, they significantly underperform. This makes the system much faster and much more efficient as we have observed in the latest Apple M1 chips. This phenomenon would make computer manufacturers more vertically integrated and therefore, the manufactured computers much cheaper, as instead of outsourcing or manufacturing external RAM units, it will be included in the chip itself resulting in less manufacturing and overall cost.

In terms of reliability, RISC architecture has fairly proven itself in markets other than Data Centers. RISC technology used by all the smartphones and tech giant Apple's switch of all their computer inventory from x86 to RISC is the sign of the system being most reliable. Additionally, reliability in terms of the computer industry is very predictable, it only takes specific benchmarks to test how reliable the system is and any fatal failure during the usage of these products is not a topic of concern. In conclusion, RISC architecture gets a green tick for Complementary Assets and reliability too.

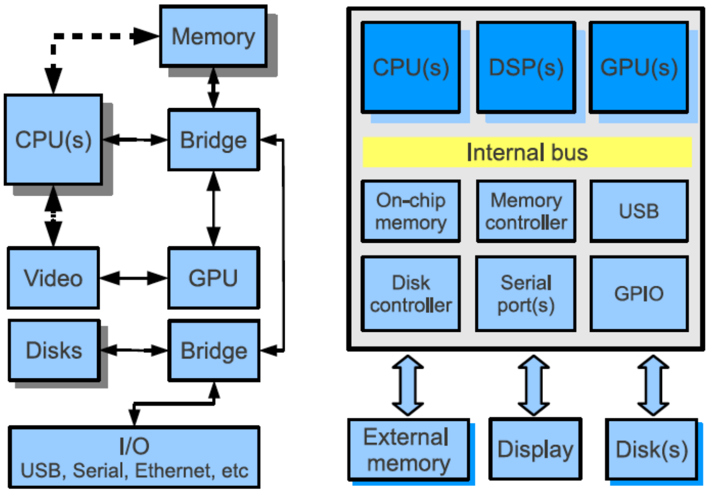


Fig 4.4 Typical x86 type system (left) and RISC based System on a chip (right)

4.3. Strategic situation in the industry

Although it was mentioned in the description of the methodology that the analysis will be held at the industry level, analyzing the strategic situation in the market of RISC architecture uncovers very interesting facts that change the whole situation in the industry. The main player in the RISC market is ARM Holdings, a chip design market that doesn't produce or sell any chips. However, the invisible contribution of ARM is more influential than almost all the giant tech companies. Such an invasion in the market has turned out because of the smartphone industry, as nearly all the smartphones are based on RISC architecture developed by ARM. However, ARM's products are not microchips, but only information. The company spends its time and resources on R&D and rather than competing with the other tech companies, they license their developments. With more than 4500 granted or pending patents, ARM is an intellectual property company. It follows a modern tech trend such as the Uber business model – a taxi company without own vehicles, Alibaba- a retailer without an inventory. This exact model can contribute to the market more than any tech company making it easier for any manufacturer to build their chips based on the designs provided by ARM. It appears to be very inconvenient for main companies in x86 such as Intel and AMD competing with this efficient business model in a long run. All in all, if some tech company decides to specialize in building computers specifically dedicated to data centers market, in terms of intellectual property infringements this tech company may not face any significant barriers which demolish entry barriers to the mentioned market which means the threat of new entrants is now a real concern for the CISC manufacturers. What this means is, ARM makes its position above the firms, to the industry level, and aims to define the whole industry. Thus, we can conclude that the strategic situation in the market is also in favor of RISC technology.

4.4. Possible Governmental Regulations and Institutional Interventions

In terms of Possible Governmental Interventions, in such global technological fields governments and institutions usually focus on environmental factors and push legislations and requirements for the firms to reduce carbon emission and power consumption. In our case, the European Union (EU) has threatened to adopt green data center legislation during the last year, to mandate data centers in Europe to be environmentally safe by 2030. However, in an attempt to circumvent government control, a group of European data center operators has joined a self-regulatory pact. 25 European data center operators and 18 industry associations have signed the Climate Neutral Data Center Pact, which establishes stringent efficiency and renewable energy goals for data centers, including a pledge to achieve climate neutrality by 2030. While the effort to reduce the impact of data centers on the environment is mostly taking place in Europe, it could have an impact in the United States. For example, many of the companies that signed the agreement in Europe are multinational, U.S. businesses, like Amazon Web Services, Google, Equinix, Digital Realty/Interxion, and CyrusOne, which are all based in the United States.

Michael Winterson, the UK managing director of Equinix, told Data Center Dynamics that the Climate Neutral Data Center Pact could be used as a good example for other things, like the EU's General Data Protection Regulation (GDPR) privacy law. People who live in the United States must follow the GDPR even if they don't live in Europe. The law also inspired California to make its strict privacy law called the California Consumer Privacy Act. For instance, the pact requires data centers to establish aggressive water conservation targets using indicators such as Water Usage Effectiveness. "If a breakthrough in water efficiency occurs in Europe, you would expect large providers to replicate it globally," Winterson said in the story.

Although the EU has not declared if the accord will put an end to their intentions for data center green laws, one EU leader told Data Center Dynamics that he supported the data center industry's effort to self-regulate. The pact highlights six areas: energy efficiency, clean energy, water efficiency, and "circular economy," which means that end-of-life IT and electrical equipment should be refurbished and reused; "circular energy," which refers to the potential for district heating systems to reuse waste heat from data centers; and governance, which requires companies to report their progress to the EU.

In the United States, legislators recently passed the Energy Act of 2020, which directs the federal government to conduct new research on the energy and water use of data centers. Additionally, it compels federal agencies to conduct energy audits and make improvements to the energy efficiency of their data centers.

From the technical comparison of the two systems, we have seen that in terms of energy efficiency RISC architecture is becoming more favorable as the systems using RISC architecture run the same tasks using more than two times less energy. On the other hand, system on a chip design allows the manufacturers to use fewer materials for manufacturing the whole logic board of the computer with all its components included at once, as the design allows the manufacturers to become vertically integrated compared to the CISC design where different components are being manufactured separately and assembled afterward. Additionally, System on a chip design allows manufacturers to further expand the system and add additional cores parallel to the system which significantly reduces the size of the system building several computers merged. Factors mentioned before are obvious from the phones and tablets that we are using. They are very fast; they fit in our pockets and don’t need any sophisticated active cooling system to keep them from overheating (Fig 4.5). Using less energy to operate, fewer materials to be manufactured, and being smaller in the size clearly for firms switching from CISC to RISC technology would be one of the first things to do to comply with the requirements of the Governments as only CPUs take roughly 37% of the total energy consumption of the Data Centers alone and adding the GPU to the system the overall consumption becomes 57% (Green Data Centers, 2007).

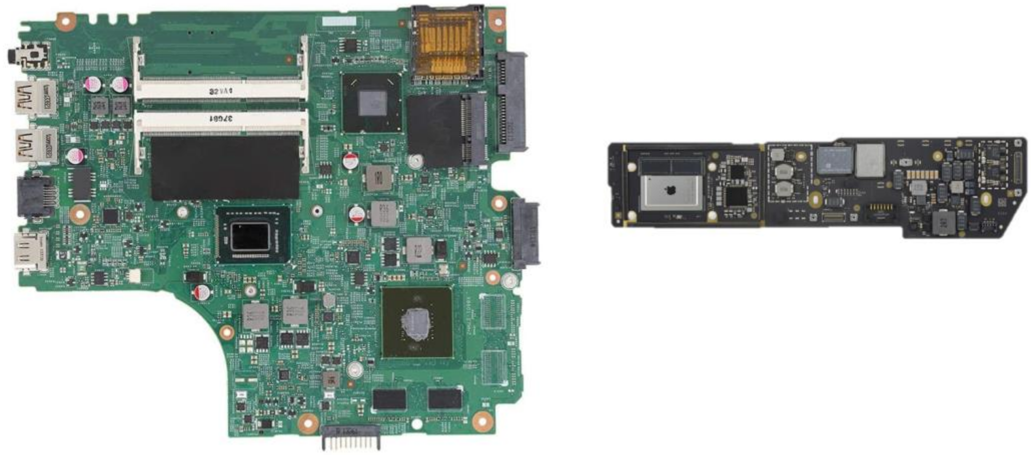


Fig 4.5 Logic board of Conventional CISC Laptop (left) vs. MacBook Air M1 RISC, roughly to scale

4.5. Switching costs

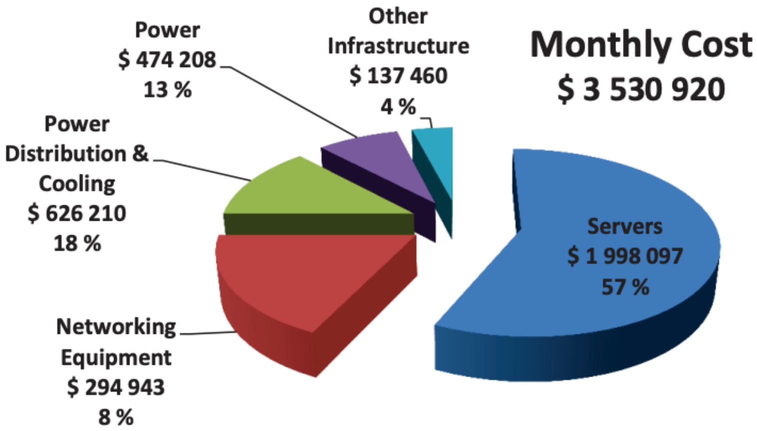


Fig 3.6 J.Hamilton's Data Center costs model

To determine the possible cost reductions associated with the use of ARM-based CPUs in a data center, the whole cost of the data center must be considered. J.Hamilton develops a cost model for a hypothetical data center and compares the costs of various components, including infrastructure, networking equipment, servers, and power. The model assumes a data center with around 50 000 servers, a ten-year amortization period for the infrastructure, a four-year amortization period for the networking equipment, and a three-year amortization period for the servers. The model estimates a 5% annual interest rate on the money utilized to fund the data center and an energy cost of $0.07 per kWh. An average critical load of 80% and a server dissipating 165 Watts is considered.

The cost savings potential of adopting ARM cortex A9 processors over the overall data center cost of roughly 10% for the Erlang SIP proxy and 12.7% for web services represented by the SPECweb2005 benchmarks is analyzed using the energy efficiency results in Tables 4.4 and 4.6 using J.Hamilton's cost model. In terms of financial cost savings, this equates to a monthly cost savings of $ 350 000 and $448 000, or a cost savings of $12,6M and $16,1M during the servers' three-year amortization period which means switching to the RISC based architecture would be beneficiary rather than carry costs in long run. The resulting monthly cost of the data center's various cost factors is depicted in Figure 4.6. It’s worth mentioning that these tests and comparisons are done using regular general-purpose RISC chips, using general-purpose logic boards. To put it in other words, if manufacturers will research and develop logic boards specifically designed for servers, savings numbers may become even more favorable for RISC architecture.

4.6. Appropriability Regime

Analyzing Regime of Appropriability leads us to the previous point about the business model of ARM Holdings which changes the structure of the whole industry. As mentioned before, ARM Holdings doesn’t own any microchips or manufacturing plants, instead, they own more than 4500 design patents and sell the intellectual property which aims at the Regime of Appropriability itself changing the whole point of it – making it easier for the newcomers to enter the market, making things complicated for the CISC industry at the other hand, opening the way for the dominance of the companies implementing RISC technology. The point is, where on one side firms are competing with each other creating and protecting their intellectual property which at some level limits the development of the technology, ARM Holdings creates ground for newcomers in the whole different technology which directly competes with the latter also contributing to the overall improvement of the technology. All in all, it can be concluded that Appropriability Regime is also in the favor of RISC design.

4.7. Analysis of the technologies from the point of view of customers

Different methodologies for customer approached analysis and comparison of the battling technologies in terms of customer requirements can be used depending on the scope of the technology itself. QFD method is considered one of the most profound methods to confront technical characteristics of technology with the exact requirements of the customers in the industry by also ranking the requirements by their importance. As a result, comparison data reveals which technology would be more beneficial for the customers.

For the simpler adaptation of QFD analysis in this Thesis work, firstly customer requirements will be taken into the account. In the case of chipset architecture in Data Servers, direct customers are Data Centers themselves who implement RISC or CISC technologies for their premises. But as there are Cloud Data Centers that are providing service to their customers which are IT companies, their needs also will be considered. For the sake of simplicity, we will select relevant requirements to assess, as although it's a very important part, the logic board architecture is only one of the many components that are present and are directly in interaction with the customers. During the assessment of the technology sector from the customers' perspective, some benchmarking and analysis results will be used and mentioned again being highlighted in the customers' point of view. The customer requirements for Data Centers considered are as follows:

- Overall Cost of the service

When it comes to Data Centers cost can be considered as one of the main aspects for the customers to decide on using the services.

The overall power consumption of a Data Center is estimated to be around 50% of the total costs of the whole premise fluctuating because of cost affecting factors depending on the location of the facility (land prices, temperature, taxes). As mentioned before, CPU and GPU together take up to 57% of the energy consumption of Data Centers including their powering and cooling. Being such a big percentage, optimization of the CPU and GPU becomes one of the main goals of the Data Centers.

In the United States, data centers consume more than 2% of total power. If they were 20% more efficient, it would take only a few years to save $2 billion nationwide.

- Available Capacity in terms of relevant Key Resources (Power, Space Cooling)

To make the best decisions when reserving space and deploying new IT equipment; utilizing power resources more efficiently; saving on operating expenses; or convincing management that the facility needs additional capacity, it is critical to have accurate and reliable real-time information about the physical space, power, and cooling in the data center. Locating and preserving resources is made much easier with real-time capacity monitoring at the site, floor, and cabinet levels.

In a Data Center Infrastructure Management system, what-if analysis helps Data Centers examine the probable net impact of planned adjustments and determine if more resources are required.

- Reliability of the system

Reliability of the system can directly be the most important factor when customers assess the Data Center they would like to choose. Within the reasonable scope, customers would choose to pay even more for better quality because the profit losses caused by unreliable systems may be devastating. Reliability is very important both in terms of power delivery and quality.

- Speed of operations

In the age of rapid technological developments, the response time has become a very delicate factor for the users. Even scrolling or searching something on the internet if a certain site loads a bit slowly, we get frustrated. End customers' satisfaction being very important for the Data Centers, slow systems cannot be accepted by IT companies as slow response times may damage the overall reputation of the company.

- Size of the system

In terms of Data Centers size directly affects the land price and overall cost of the system. A room full of servers takes quite a big space depending on the capacity of the premise. The bigger the facility, the more complicated the cooling, higher the land price. All in all, size can directly affect the cost of the premise.

- Repairability / Maintenance of the system

Repairability and Maintenance of the system, in the long run, can become a price-affecting factor in the Data Centers. Overall, customers are more prone to being able to repair the goods they are receiving. In terms of Data Centers high repairability would allow changing vital, most vulnerable parts of the system such as GPU, CPU, RAM if they fail during their operational life.

In the second phase, we will select the technical characteristics of the technology that affects these customer requirements. Those are:

- Power Efficiency

Power efficiency is one of the key parameters affecting the overall cost of the system has been compared in 4.1. Technological Comparison of the Methodology chapter. For the sake of simplicity, power efficiency will cover all the technical benchmarks in its name. From the attained results of the two different benchmarks, it's been concluded that RISC architecture is superior in comparison with CISC showing better results on almost all the benchmarks demonstrating much better Requests per Joule, Number of simultaneous sessions per Joule, and Number of calls per Joule. Additionally, testing on HTTP and SQL servers, RISC-based devices ended up being between three and four times more energy efficient compared to CISC systems when the requests/s/watt ratio is considered under various loads.

- Physical Dimensions

Physical dimensions play a key role in reducing the overall cost of the Data Centers as smaller systems can be fitted into smaller rooms and consequently, lead to lower land costs and lower cooling costs as it is cheaper to maintain air conditioning and cooling of a smaller system. Additionally, usage of fewer materials for the facility and equipment would lead to cost savings. In terms of RISC and CISC comparison, in 3.4. it's been mentioned that the RISC architecture allows the system to become much smaller and have better overall performance than CISC. On the other hand, because of the overall high heat exchange during the operation of CISC computers sophisticated and chunkier heatsinks and fans are used. Some premises even use liquid cooling solutions to maintain the system cooler which makes the design even bigger. On Fig. 3.5. we can see computers with two different architectures compared with each other in terms of size. MacBook M1 Air logic board on the right picture works so cool that the manufacturer even didn't include any active cooling system in the computer to keep it cool whereas, even on very low budget CISC computers active cooling is required. Finally, as was predicted, in the latest event of Apple, the new chip that the manufacturers revealed merges 2 M1 chips together which is only possible in RISC architecture. It is predicted that with several hundreds of server-in-a-chip attaining thousands of cores on a single board is a possibility that is not far from now on (S. Saponara, L. Fanucci, M. Coppola 2012). If this will be done, very small-sized servers will become a reality.

- Purchase and maintenance costs

At 3.5. Switching costs, the direct comparison between CISC and RISC costs have been studied as it describes the cost of switching from CISC to RISC. The results have concluded that even in this stage it would be more efficient to switch to RISC. If the server-in-a-chip design will be a reality, the purchase costs of RISC will become much less than CISC. Additionally, the cooling system is considered the most fragile component of computers and if not maintained properly, they tend to lose their efficiency over time. Having more sophisticated cooling systems would make the maintenance costs of CISC more than RISC architecture. On the other hand, RISC architecture merging most vital components in one chip means if one of the components end its life, the whole chip should be replaced which may induce more costs.

- Life expectancy

Life expectancy is one of the technical characteristics that directly affect the overall costs and reliability of the system. In terms of RISC and CISC, they both are quite reliable with the same life expectancy for both systems, because they both are made of Silicon and when considering the life of electronic chips, usually the life of silicon is considered which is 7-10 years.

- Expandability

Expandability describes the easiness of adding new modules to the existing Data Center when required expanding the needed power delivery, space, and cooling. As been discussed before, it is more efficient to expand RISC systems physically and it would take less cost as added modules need less power to operate in contrast to CISC, take less space, and need a less sophisticated cooling system.

Reporting the Customer requirements by their importance and comparing the technical characteristics, the importance factor is reported on a scale of 1-10. On the other hand, for the comparison of technical characteristics usually 1, 3, 9 rankings are used, and a better system gets a higher score. For the sake of simplicity, as in this industry, there are only 2 competing technologies are present, the results will be taken into the consideration roughly, not by the exact data results, but by the winner and loser at each technical characteristic. We will also skip the strength of the relationship between technical characteristics and Customer Requirements. Results of the analysis are as follows:

Customer Requirements	**Importance Factor**
Cost of the Service	9
Available Capacity	6
Reliability	10
Speed of operations	9
Size of the system	8
Repairability	6

Technical Characteristics	**RISC**	**CISC**
Power Efficiency	9	3
Physical Dimensions	9	3
Purchase and maintenance costs	9	9
Life expectancy	9	9
Expandability	9	3

As the next step to compare the two technologies weighting the scores of the technical characteristics by the importance is required. To do this all the relevant technical characteristics will be summed and multiplied by the given importance factor and reported in the new table that directly compares RISC and CISC from the point of view of customers:

- Cost of the service:

  RISC: 9*(Power Efficiency (9) + Physical Dimensions (9) + Purchase and maintenance costs (9) + Life expectancy (9)) = 324

  CISC: 9*(3+3+9+9) = 216
- Available Capacity:

RISC: 6*(Power Efficiency (9) + Physical Dimensions (9) + Expandability (9)) = 162

CISC: 6*(3+3+3) = 54

- Reliability

  RISC: 10*(Life expectency (9) + Purchase and maintenance costs (9)) = 180

  CISC: 10*(9+9)=180

- Speed of Operations

  RISC: 9* Power Efficiency(9) = 81

  CISC: 9* 3 = 27

- Size of the system

  RISC: 8*(Physical Dimensions (9) + Purchase and maintenance costs (9) + Expandability (9)) = 216

  CISC: 8*(3+9+3) = 120

- Repairability

  RISC: 6* Purchase and maintenance costs (9) = 54

  CISC: 6*9 = 54

Customer Requirements	**RISC**	**CISC**
Cost of the Service	324	216
Available Capacity	162	54
Reliability	180	180
Speed of operations	81	27
Size of the system	216	120
Repairability	54	54
Total Score	1017	651
Normalized score	**1.00**	**0.64**

5. Conclusion

Data centers are one of the most vital components of modern IT companies and companies requiring database or additional computing power for their operations. At first, companies that require a data center had their own facilities, however this trend is changing for the past decade. Companies slowly move to cloud premises which moves traditional on-site data centers to off-site, which can save the company from huge investments for building a facility instead by leasing the service. This vertical disintegration trend makes individual stand-alone data center companies to work more on their efficiency to reduce the costs of the service and improve their technology also because of a competition in the individual data center market. The core and the most power-hungry component of data centers is CPU. Consequently, in the industry mostly focus for the improvement is on CPU. Currently Complex Instruction Set Computer (CISC) architecture is the dominant design which is used in almost all Data Centers. This architecture was created at the first times when computers were starting to develop, and Intel was the pioneer of this technology. The architecture was created to give more weight to the hardware instead of software for computing as back then programmers didn't have sufficient experience and knowledge. Reduced Instruction Set Computer (RISC) architecture, however, gives more weight on the software making the programming process harder, on the other hand, making the operation of the system more efficient. In theory, during the long-term use it obvious that RISC would be more energy efficient. Currently, regardless of the yearly improvements, improvement of CISC technology stalls because of its design limitations. In this research paper, these two battling Instruction Set Architectures in Data Centers were analyzed. CISC currently being the dominant design in the industry has recently been facing challenges by RISC architecture about which there are scattered information and predictions mainly stating that soon RISC will dominate the market. To make more precise assumptions an analysis taking the most important factors that are capable change the way of the industry was made. The analysis is based on F. Suarez's integrative framework for focusing on the right factors affecting the dominance in the market of a firm on right times. However, the framework developed in this thesis focuses more on the technology itself from the industry level and the analysis focuses more on the environment where the technology is introduced. To achieve better results every possible aspect that may change the industry within limitations was taken into the consideration. The framework also assumes that the market is already established and there is/are dominant design in the industry. History shows that for a certain innovation to become dominant in the market by replacing the already existing dominant design the new design should have a technological advantage over the old one. Having met this condition, the next factors were assessed which can change the direction of the industry. During the analysis, both technologies were analyzed from the industry level, and advantage or disadvantage of the proposed technology is assessed. The considerations were as follows:

1. Technological Comparison
2. Complementary assets
3. Strategic situation of the industry
4. Possible Governmental Regulations and institutional interventions
5. Possible Switching costs
6. Appropriability regime
7. Customers' point of view

It is assumed that the more weight the new technology has in these considerations, the more chances to become a dominant one it has. In the case of the selected technology, the report plotted on Fig 5.1 as the result of analysis demonstrates that RISC technology has an absolute weight over CISC and has all conditions to become a dominant design in the market in near future and it can be predicted that if the right steps towards developing this technology further will be made, it has almost a certain chance to become a dominant design in the Data Centers market. The analysis in terms of technological comparison showed that although in many categories RISC dominates, there may be some operations such as floating-point calculations where CISC is still better than RISC. This in long-term might mean CISC may stay in the market by working in a hybrid manner with RISC being dominant in the market soon.

The framework that has been developed for this research can also be used for different technologies and markets to predict the state of the battle between several designs by adapting the framework to them. Fig 5.1 depicts what this framework looks like for the example made in this thesis. By researching the industry and the market according to these criteria it is possible to have a more detailed approximation about the considered technology.

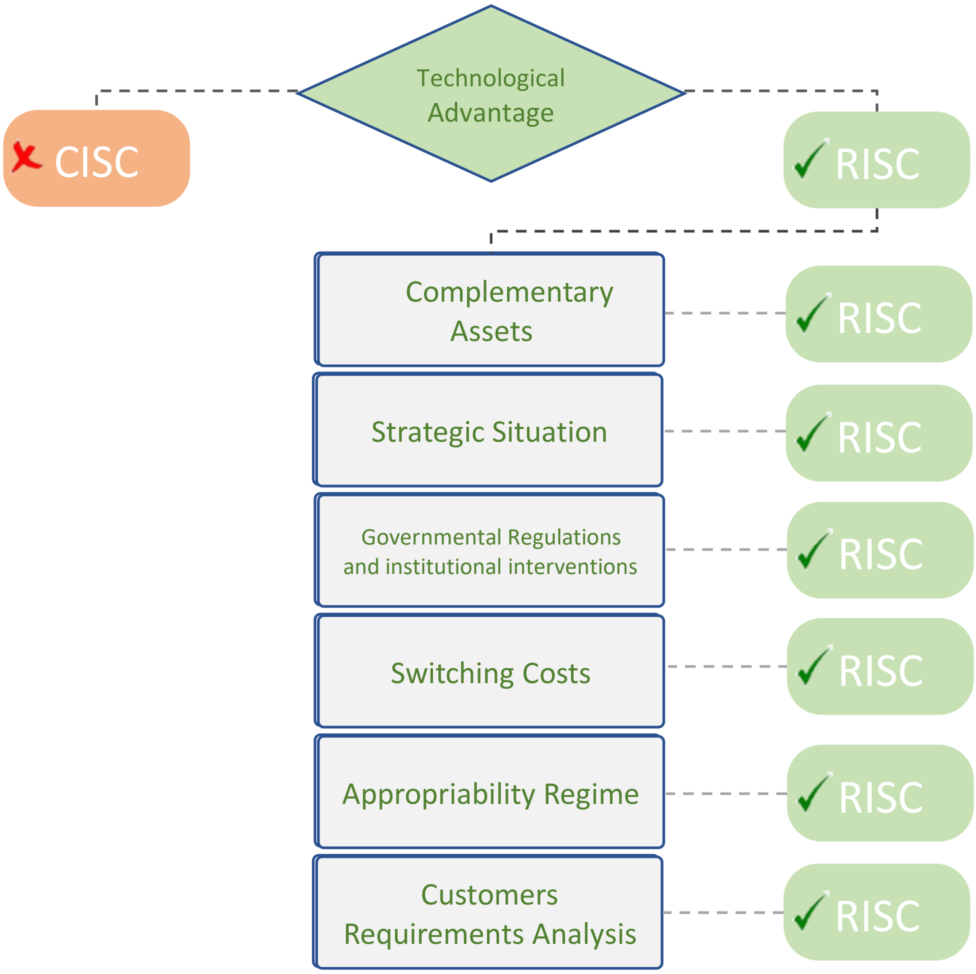


Figure 5.1 Framework for prediction of the success chances of RISC architecture in Data Centers.

Sitography

www.britannica.com

www.computerhistory.org

www.ibm.com

www.omnisci.com

www.nvidia.com

www.intel.com

www.javapoint.com

www.paessler.com

www.arm.com

www.microcontrollertips.com

www.networkworld.com

www.paloaltonetworks.com

www.perspectives.mvdriona.com/2010/09/18/OverallDataCenterCosts.aspx

www.omnisci.com/technical-glossary/cpu-vs-gpu

www.extremetech.com

www.somlabs.com

www.tomshardware.com

www.extremetech.com

www.statista.com

www.sixsigmastudyguide.com

References

- Schilling, Melissa (1998), “Technological Lockout: An Integrative Model of the Economic and Strategic Factors Driving Technology Success and Failure,” Academy of Management Review, 23 (2), 267–84.

- Raji Srinivasan, Gary L. Lilien, & Arvind Rangaswamy (2006) “The Emergence of Dominant Designs”.

- W.J. Abernathy, J.M. Utterback, “A Dynamic Model of Process and Product Innovation”, 1975.

- W.J. Abernathy, J.M. Utterback, “Patterns of Industrial Evolution”, 1978.

- R. Srinivasan, G.L. Lilien, A. Rangaswany, “The Emergence of Dominant Designs”, Journal of Marketing, 2006.

- W. J. Abernathy, “Productivity Dilemma: Roadblock to innovation in the Automobile Industry” 1978, Baltimore.

- Katz, M., Shapiro, C., 1986. “Technology adoption in the presence of network externalities”. Journal of Political Economy 94, 822–841.

- Teece, D., 1986. “Profiting from technological innovation: Implications for integration, collaboration, licensing, and public policy”. Research Policy 15, 285–305.

- F.F. Suarez, 2003. “Battles for technological dominance: an integrative framework”, London Business School, pp. 274-279.

- Garud, R., Rappa, M., 1995. “On the persistence of researchers in technological development".

- Luiz G, Rafael A., 2012 “Towards green data centers: A comparison of x86 and ARM architectures power efficiency”.

- Robert Basmadjian, Pascal Bouvry, Georges da Costa, László Gyarmati, Dzmitry Kliazovich, et al.. Green Data Centers. Large-Scale Distributed Systems and Energy Efficiency, Wiley, pp.159-196, 2015, ff10.1002/9781118981122.ch6ff. ffhal-01196827f.

- S. Saponara, L. Fanucci, M. Coppola, "Many-core platform with Noc interconnect for low cost and energy sustainable cloud server-on-chip" Sustainable Internet and ICT for Sustainability (SustainIT), 2012, 2012, pp. 1–5.

- A. Akiike., 2013. "Where is Abernathy and Utterback Model?". Annals of Business Administrative Science 12, 225-236.

- Anderson, Philip and Michael L. Tushman (1990), "Technological Discontinuities and Dominant Designs: A Cyclical Model of Technological Change," Administrative Science Quarterly, 35 (4), 604–633.

- Besen, Stanley M. and Joseph Farrell (1994), "Choosing How to Compete: Strategies and Tactics in Standardization," Journal of Economic Perspectives, 8 (2), 117–31.

- Katz, M. and Carl Shapiro (1986), "Technology Adoption in the Presence of Network Externalities," Journal of Political Economy, 94 (4), 822–41.

- Schilling, Melissa (1998), "Technological Lockout: An Integrative Model of the Economic and Strategic Factors Driving Technology Success and Failure," Academy of Management Review, 23 (2), 267–84.

- James M. Utterback (1995), "Dominant Designs and the Survival of Firms," Strategic Management Journal, 16 (6), 415–30.

- Fernando F. Suar z, and James M. Utterback (1998), "Strategies for Survival in Fast-Changing Industries," Management Science, 44 (December), 207–220.

- J.M Utterback, "Mastering the Dynamics of Innovation", Harvard Business School Press, 1996.