Embedded Systems
Theory

Christian Rossi

Academic Year 2024-2025

Abstract

The course delves into Embedded Systems, covering their characteristics, requirements, and constraints. It explores hardware architectures, including various types of software executors, communication methods, interfacing techniques, off-the-shelf components, and architectures suited for both prototyping and large-scale production.

In terms of software architectures, the course examines abstraction levels, real-time operating systems, complex networked systems, and the tools and methodologies used for code analysis, profiling, and optimization.

Students will also learn to analyze and optimize hardware/software architectures for embedded systems, focusing on managing design constraints and selecting appropriate architectures. Key topics include estimating and optimizing performance and power at various abstraction levels, project management, and designing for reuse.

Additionally, the course addresses run-time resource management and includes case studies to illustrate trade-offs based on application fields and system sizes.

Contents

CHAPTER 1

Introduction

1.1 Introduction

Embedded systems are characterized by their ubiquitous presence, low power consumption, high performance, and interconnected nature. These systems are commonly utilized in four main application contexts:

- *Public infrastructures*: safety is a primary concern to prevent potential attacks, and in some cases, latency is also crucial. Examples include highways, bridges, and airports.
- *Industrial systems*: reliability and safety are the predominant concerns. Key industries utilizing embedded systems include automotive, aerospace, and medical.
- *Private spaces*: this includes control systems for houses and offices.
- *Nomadic system*: these systems involve data collection related to the health and positions of animals and people, requiring both security and low latency.

In the future, embedded systems will evolve in several key ways:

- *Networked*: transitioning from isolated operations to interconnected, distributed solutions.
- *Secure*: addressing significant security challenges that impact both technical and economic viability.
- *Complex*: enhanced by advancements in nanotechnology and communication technologies.
- *Low power*: utilizing energy scavenging methods.
- *Thermal and power control*: implementing runtime resource management.

Usually, we process useful data locally and send the relevant data to the cloud less frequently. This approach conserves energy, thereby extending battery life. Due to time constraints, developing a device from scratch is often infeasible, so we typically collaborate with eco-alliance partners. The design process involves a multidisciplinary team, as it integrates multiple domains of expertise. The first prototype resulting from this design phase is called the Minimum Viable Product (MVP), which is the initial sellable version of the product.

1.1.1 Technological problems

Applications are expanding rapidly, pushing the need for mass-market compatibility and integrating into all aspects of life, which in turn drives up volumes. As technology evolves, the scale of integration grows, supported by new materials and programming paradigms. However, finding a balance in this progress is complicated by several factors. CMOS technology is reaching its physical limits, and the cost of developing new foundational technologies can often be unaffordable. Additionally, the power and energy demands create significant barriers, as does the exponential proliferation of data. Transitioning from invention to innovation is proving difficult, and the design methodologies in use today heavily exploit human effort—though this reliance on human ingenuity is, in some ways, a fortunate necessity.

Vertical applications Vertical applications requires various technologies to work ad desired such as sensors, computing units, storage, communication elments, and so on. Energy and power dissipation have become even more problematic with the introduction of the newest technology nodes, exacerbating existing challenges. Dependability, which encompasses security, safety, and privacy, is now a major concern. At the same time, the growing complexity of systems is reaching nearly unmanageable levels. Despite this, complexity continues to rise, driven by applications that build upon increasingly interconnected systems of systems.

1.2 Applications future

Future applications will be highly compute-intensive, requiring efficient hardware and software components across various domains, including embedded systems, mobile devices, and data centers. Key characteristics of these future applications include:

- *Compute-intensiveness*: they will demand significant computational resources, necessitating optimized hardware and software regardless of their application domain.
- *Connectivity*: these applications will be interconnected, either wired or wirelessly, and will often be online—globally interconnected through the Internet.
- *Physical entanglement*: they will be embedded within and capable of interacting with the physical world, not only observing but also controlling their environment. These systems will effectively merge into our everyday surroundings.
- *Intelligence*: these applications will possess the capability to interpret noisy, incomplete, analog, or remote data from the physical world, allowing for smarter interactions and decision-making.

All major future applications will exhibit these traits to varying degrees. The ongoing integration of the digital and physical worlds (manifested through the Internet of Things (IoT) and Cyber-Physical Systems (CPS)) will be driven by advances in cognitive computing, big data analytics, and data mining.

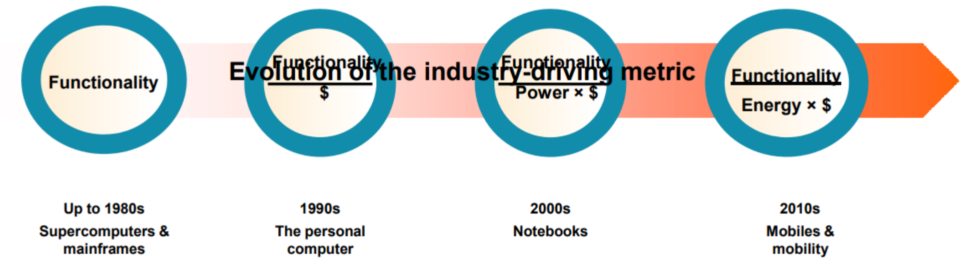

Figure 1.1: Industry requirements

The challenge is that optimizing for functionality, energy, and cost remains difficult:

$$\frac{\text{functionality}}{\text{energy} \times \text{cost}}$$

However, successfully optimizing this metric often leads to improvements across simpler metrics as well.

While the number of transistors in modern systems can continue to increase, a key challenge is that we cannot power all of them simultaneously. This limitation drives the need for innovative approaches to using extra transistors more efficiently. Techniques like multi-core and many-core processors, as well as domain-specific processors, are becoming essential. Additionally, heterogeneous processing combined with aggressive power management is crucial for optimizing performance across various tasks. As data generation continues to accelerate it becomes increasingly important to ensure that computation is carried out in the most efficient location. This efficiency is necessary to cope with the imbalance between the massive growth of data and the slowing progress of Moore's Law.

1.2.1 Transistors and cores

As transistors shrink in size, several challenges emerge. Process variation, physical failures, and aging mechanisms, such as negative bias temperature instability (NBTI), can degrade device performance over time. With very-large-scale integration (VLSI), packing more transistors into smaller spaces increases power density, which in turn creates thermal issues. Furthermore, traditional communication subsystems struggle to provide adequate power-performance trade-offs in these densely packed systems.

For instance, Network-on-Chip (NoC) designs can consume up to 30% of total chip power, and their performance plays a critical role in the efficiency of multicore architectures, which are increasingly necessary for improving system performance across various applications.

1.2.2 Silicon challenges

Despite silicon being a relatively good thermal conductor (about four times worse than copper), large chips can still experience significant temperature gradients, especially in high-performance CPUs. This creates hot spots that can affect chip reliability and performance. One practical solution involves dividing the chip into concentric rings and applying dynamic voltage and frequency scaling (DVFS) to each region. By fine-tuning voltage and frequency for each ring, it becomes possible to optimize performance while maintaining thermal balance, preventing excessive heat buildup in certain areas.

1.2.3 Frequency scaling

Modern systems address power management by partitioning chips into independent islands that operate at different voltage and frequency levels. This approach allows sections of the chip to be dynamically turned off when not in use, a process known as power gating. By adjusting power usage in this way, overall system energy efficiency is greatly improved without compromising performance when it is needed.

1.2.4 Cooling systems

Cooling systems are critical to managing the thermal output of modern computing hardware, and various methods are used depending on the specific requirements. Air cooling, while common, suffers from a low heat transfer coefficient (HTC), poor chip temperature uniformity, and requires large heat sinks and air ducts, particularly in data centers. It is also noisy and expensive to maintain. Water cooling offers an improvement, with better HTC, more uniform chip temperatures, smaller heat sinks, and fewer fans. Water cooling also allows for potential heat recovery, though it requires large pumps to function effectively.

Two-phase cooling systems present an even more efficient solution. They provide higher HTC, better chip temperature uniformity, and smaller pumps, along with isothermal coolant, which helps maintain consistent temperatures across the chip. This method excels at cooling hot spots and also allows for heat recovery. However, two-phase cooling systems suffer from low pump efficiency and reliability issues.

New innovations, such as thermosyphon cooling, are emerging as promising alternatives, offering advanced methods to maintain chip performance while effectively managing heat dissipation.

1.3 Productivity and planning

The revenue model can be visualized simply as a triangle, where the product's life is represented by a span of 2W, peaking at W. The time of market entry defines the triangle, representing market penetration, with the area of the triangle corresponding to total revenue. Any delay in market entry results in a loss, which is the difference between the areas of the on-time and delayed triangles.

The productivity gap, however, reveals a more challenging situation. In theory, increasing the number of designers on a team should reduce project completion time. In practice, though, productivity per designer tends to decrease due to the complexities of team management and communication, a phenomenon famously referred to as the mythical man-month (Brooks 1975). At a certain point, adding more designers can even extend project timelines, a classic case of too many cooks spoiling the broth.

1.3.1 Platform-Based Design

To enhance productivity, the design methodology must support reuse, particularly at higher abstraction levels, and this should be backed by standardization. As integrated systems increasingly require both digital and non-digital functionalities, this dual trend is captured by the International Technology Roadmap for Semiconductors. It emphasizes the miniaturization of digital functions, often referred to as More Moore, and functional diversification, known as More-than-Moore.

1.3.2 Technology Trends

Two significant trends in technology development are System-on-Chip (SoC) and System-in-Package (SiP). SoC focuses on full integration and achieving the lowest cost per transistor, while SiP focuses on lowering the cost per function for the entire system. These architectures are complementary rather than competitive, each requiring distinct industrial approaches and advanced research and design knowledge. SoC emphasizes miniaturization, whereas SiP centers on integrating multiple components, necessitating different manufacturing competencies for each.

MEMS Micro-Electro-Mechanical Systems (MEMS) involve the creation of 3D structures using integrated circuit fabrication technologies and specialized micromachining processes, typically on silicon or glass wafers. MEMS devices include transducers, microsensors, microactuators, and other mechanically functional microstructures. Applications range from microfluidics (valves, pumps, and flow channels) to microengines (gears, turbines, combustion engines). Integrated microsystems combine circuitry and transducers to perform tasks autonomously or with the assistance of a host computer. MEMS components bridge the gap between the electrical and non-electrical world, where sensors receive inputs from non-electronic events and actuators output to them.

Energy Scavenging Energy scavenging involves capturing energy from objects with temperature gradients. Another source of scavenging is vibrations, such as self-winding watches, which generate around 5 microwatts on average when worn and up to 1 milliwatt when shaken vigorously.

Wireless Sensor Nodes Wireless sensor nodes are small, battery-powered devices that monitor local conditions. These devices typically have limited resources and form nodes within a wireless network that covers a region or object of interest. Wireless sensor nodes enable new applications by collecting, fusing, reasoning, and responding to sensor data. These applications can lead to smarter systems in fields ranging from environmental monitoring to industrial automation.