Why do we need to focus on SLMs?

*By Subhra:* [*https://www.linkedin.com/in/sahoosubhra/*](https://www.linkedin.com/in/sahoosubhra/)

**The question I am trying to answer, in this article is- “Will knowledge distillation, combined with quantization, become the most prevalent and effective method for creating high-performing SLMs, thereby accounting for most commercial applications (Inference) within the next three years?”**

[Beyond the Sum of its Parts 1](#_Toc81225949)

[A simple bottleneck scenario 1](#_Toc1403411280)

[Less Performant Processors 1](#_Toc662247615)

[Expensive Datacenters 1](#_Toc1319382537)

[The trend of On-device processing 1](#_Toc906088865)

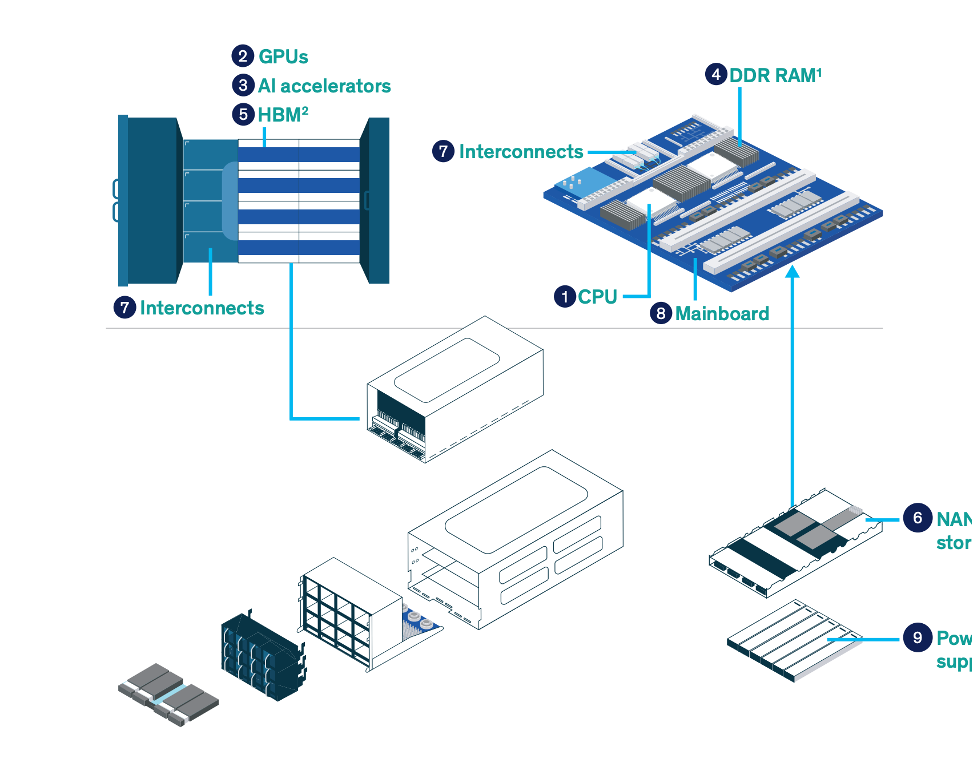
[What is SLM? 1](#_Toc429607259)

[Strategic Question 1](#_Toc1297935744)

[Applications 1](#_Toc2058991466)

The physical bottle necks in LLMs are helping the growth of SLM. In this article I explore the strategic question:

# Beyond the Sum of its Parts



A modern artificial intelligence server, constituents – the central processing unit (CPU), graphics processing units (GPUs), sophisticated storage solutions, high-throughput networking, and robust power delivery systems – each possess distinct logistical and, critically, economic attributes that shape deployment and scalability.

**Processing**

CPUs and GPUs execute the intricate instructions underpinning AI, with specializzed AI accelerators emerging as crucial adjuncts, their performance often benchmarked in FLOPS (Floating Point Operations Per Second). The relentless pursuit of greater efficiency and speed is intrinsically linked to advancements in semiconductor fabrication, notably the shrinking transistor node size and the characteristics of the silicon wafer itself (thickness), factors that directly impact both performance and cost.

**Power & Infrastructure**

Rack density, a key metric in data center economics, dictates space utilization and cooling requirements. A stable and scalable power supply is non-negotiable for continuous operation, while the mainboard acts as the critical nexus, its design influences both connectivity and overall system efficiency.

**Storage/Memory**

NAND flash storage offers high-capacity, persistent data holding, while DDR RAM provides the rapid access necessary for active processing. High Bandwidth Memory (HBM) represents a premium solution for accelerating data-intensive tasks. The interplay between the speed, capacity, and cost of these storage and memory tiers significantly impacts the economic viability of AI deployments.

**Interconnects**

Finally, the efficiency of an AI server hinges on the interconnects that facilitate data flow between its various components. These pathways are crucial for ensuring seamless communication and maximizing the collaborative potential of the system. Bottlenecks here can undermine the performance gains achieved in individual processing units, highlighting the importance of holistic system design.

|  |  |  |
| --- | --- | --- |
| **Category** | **Component/Factor** | **Description** |
| Processing | CPU | Determines how fast tasks are done. |
|  | GPUs | Specialized processors that significantly speed up tasks like AI and complex calculations by doing many things at once. |
|  | AI Accelerators | Hardware specifically designed to make AI tasks run much faster and more efficiently. |
|  | Transistor Node | The size of the tiny switches inside chips; smaller means more power in less space, leading to better performance and efficiency. |
|  | Wafer Thickness | The depth of the silicon material used to make chips; it can affect how well the chip handles heat and its physical durability during manufacturing and use. |
| Power & Infrastructure | Rack Density | How many computing units can fit into a standard equipment space; higher density needs careful power and cooling management. |
|  | Power Supply | The component that provides electricity to the system needs to be strong enough for all the parts. |
|  | Mainboard | The main circuit board connects all the different parts of the computer and distributes power. |
| Storage/Memory | Storage Capacity | The total amount of digital information the system can hold for later use and for active work. |
|  | NAND Storage | Long-term, non-volatile storage like a digital filing cabinet for large amounts of data (like in SSDs). |
|  | DDR RAM | High-speed "working memory" that the computer uses for tasks it's currently doing; more RAM allows for smoother multitasking with large data. |
|  | HBM | Very fast memory used with powerful processors like GPUs and AI accelerators to quickly access large amounts of data needed for intensive tasks. |
| Other | Interconnects | The pathways and technologies that allow different components to communicate and share data within the system. Faster and more efficient pathways improve overall speed. |

# 

# A simple bottleneck scenario

## Less Performant Processors

|  |  |
| --- | --- |
| **Processing**  The performance increase in processors is not proportional to increase in power consumption | Inserting image... |

## Expensive Datacenters

|  |  |
| --- | --- |
| **Power & Infrastructure**  As a result, datacenters are becoming expensive to operate | Inserting image... |

# The trend of On-device processing

The intense computational and energy demands of large language models (LLMs) are creating significant economic pressures, driving strategic investments in areas from graphene processors to nuclear energy. Software innovations aimed at efficiency are also accelerating, as evidenced by use cases from firms like Apple and Lux Capital. A key trend is the rise of small language models (SLMs), by major AI vendors and adopted by companies like Apple, as a compelling alternative that addresses power consumption, hardware limitations, and data privacy concerns by enabling on-device processing. This economic reality is thus shaping the future of AI development and deployment.

Small Language Models offer a compelling alternative

* Major AI vendors promoting SLMs (Microsoft, Google, OpenAI).
* Apple adapting to SLMs.
* SLMs address power consumption and hardware limitations.
* Ensure data privacy as they run within our infrastructure.

## What is SLM?

Small language models (SLMs) are artificial intelligence (AI) models capable of processing, understanding and generating natural language content. As their name implies, SLMs are smaller in scale and scope than large language models (LLMs).[1]

LLMs serve as the base for SLMs. Model compression techniques (Pruning, Quantization, Low-rank factorization, Knowledge distillation) are applied to build a leaner model from a larger one. Compressing a model entail reducing its size while still retaining as much of its accuracy as possible. [1]

Examples of SLMs: DistilBERT, Gemma, GPT-4o mini, Granite, Llama, Mistral, Phi

# Strategic Question

|  |  |  |  |
| --- | --- | --- | --- |
| **Will knowledge distillation, combined with quantization, become the most prevalent and effective method for creating high-performing SLMs, thereby accounting for most commercial applications (Inference) within the next three years?** | | | |
| **Scenarios:** | **Why Yes?** | **Why No**  **(i.e. LLMs are better)?** | **What if -**  **GPU Alternative?** |
|  |
| **Logistics** |  |  |  |
| **Physics** | Low end Infra, good performance on AI Accelerators | High-end Infra | example ASIC |
| **Dominant Design** | Not yet, but they follow the form and factors of LLM. RAG, Assistants but not Agents | LLMS can be used for Assistants, Agents or RAG Applications |  |
| **Production Technology** | Tools are the same as LLMS, Inference in cheaper. Finetuning is cheaper | Expensive Infrastructure, Expensive to retrain | *Unknown* |
| **Technological Bottlenecks** | GPU are required to finetune, or achieve an acceptable response time | High end GPUs are required | *Unknown* |
|  | | | |
| **Economics** |  |  |  |
| **Cost** | Inference/deployment costs are significantly reduced. But internal effort for Provisioning | Higher Cost (License, Subscription) | *Unknown* |
| **Size** | Smaller Models | Larger Models | Larger Models, but SLMs will run on these too |
| **Power Consumption** | Lower power consumption | High power consumption | *Unknown* |
| **Speed** | Low response time, Effort invested on Development, Provisioning etc. | Low response time | *Unknown* |
| **Accuracy** | Good Enough | High | *Unknown* |

# Applications