Edge AI constraints and current limitations

Paolo Azzoni – INSIDE Industry Association

05/12/2024

AI Pillars

NEW ALGORITHMS

Recent advances have established **a new approach to statistical computing**, distinct from traditional "expert systems" that rely on algorithms and software engineering for precise calculations. E.g. neural-oriented computing.

DATA AVAILABILITY

Neural computing is data-driven: no explicitly programmed rules, example-based training allowing to approximate solutions for new, unseen inputs. The effectiveness of these results heavily depends on **large structured and semi-structured datasets, which are today available**.

COMPUTING POWER

The ability to experiment with new algorithms on large datasets requires **powerful computing infrastructures**. Without them it would be impossible to handle the demand of **training** on vast datasets, running **inferences** on new data, and **validate** the effectiveness of AI.

These are the three main enablers of modern AI.

Modern AI constraints

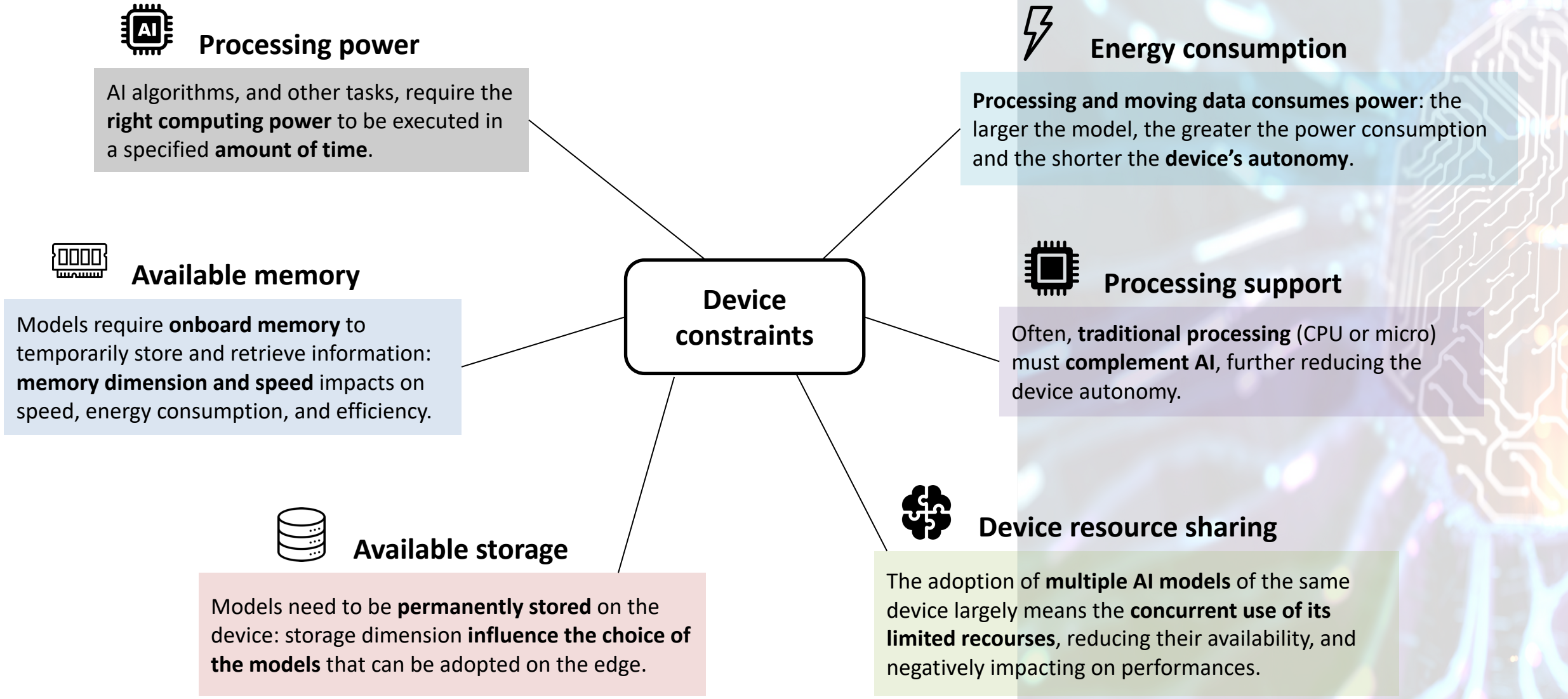
What are the constraints of modern AI?

1. Constraints related to device resources, features & capabilities.
2. AI model and application constraints.
3. Environmental and financial constraints.

Current AI stack (HW&SW) based on semiconductors has also several limitations.

Constraints & limitations become more critical for embedded intelligence (edge AI).

Edge device constraints



Edge model & application constraints

Model & application constraints

Model size

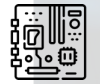
Large models typically necessitate increased **computational power and memory**, often resulting in slower operation.

Application defines speed

Edge devices, are resource constrained, and **could not be capable to run the AI models** (ingest data and run inference) at the **speed** required by the **application**.

Model accuracy/precision

The **precision** adopted to represent data and models affects the necessary **hardware resources**, thus also the accuracy of AI and its performances.



Data vs resource vs application

The use of **large datasets or high-resolution** data can quickly and largely **exceed** the available device resources, **preventing their use in the application**.

Architecture of the model

The **structure and interconnection of parameters** within a neural network affect its **computational efficiency, memory utilization and operational speed**.

Raw data preprocessing

In addition to AI, very often, **raw data pre-processing** before ingesting data in the model requires extensive computing and memory resources.

Environmental, operating & financial constraints

Device form factor

Edge devices need to satisfy specific dimensions and weight requirements: components like cooling, interfaces and batteries make it difficult to satisfy them.

Environmental aspects

Edge devices must often provide **extended operating environmental conditions** (e.g. extreme temperatures, humidity, dust, radiation), requiring **high reliability HW**, which is generally less performant.

Safety & security

Environments involving safety & security often require **hardware redundancy**, which impacts on other constraints and on costs.

Environmental, operating & financial constraints

Accessibility

Accessing devices on the edge is often **difficult and expensive, impossible** in certain applications.

Deployment & commissioning

Deployment and commissioning of device on the edge is often difficult and expensive!

Maintenance & evolution

Operating, managing, updating, maintaining, replacing, retiring and recycling devices on the edge is expensive!

Computing limitations

Acceleration is mostly based on brute force (smaller transistors, more transistors in chips, more cores, more operations per seconds, etc.) and architectures improvements

- **Transistor miniaturisation** has physical limits (Moore's law):
  - Thermodynamics (experiments reached 0.34 nm …)
  - Manufacturing process limitations (e.g. photolithography)
- **Options to address limitations:**
  - 3D technologies and heterogeneous integration.
  - Special purpose chips (Nvidia, Google, Amazon, Tesla, etc.)
  - Find a transistor replacement
    - Use light (Integrated photonics: 20 times faster than electrons; limitation is density)
    - Memristors (a resistor with memory, retaining resistance changes; scale down to 1 nm)

Computing limitations (2)

- **Memory wall**: a significant portion of time is spent accessing the memory and memory bandwidth doesn't scale at the required pace.
  - A model that today requires 6 months of training, using 60% more the GPU, requires about 2.5 months just for transferring data to and from the memory.

  Solutions:

  - 3D technologies: to increase the width of the memory bus.
  - Increase clock speed (not significantly increased in the last decade).
- **Energy limitations**: energy inefficiency is becoming a major barrier to sustainably scaling AI systems (human brain consumes around 20 Watts also in most demanding computations …).

The technology stack is a must

Efficient hardware is a must! But unless it offers clear physical superiority, it is not enough for competitiveness.

- Hyperscalers strategy doesn't consist **only of building chips**: chips don't differ substantially in processing bits …
- **The differentiating element is the capability to control the entire stack!**
- This approach called "**verticalization**" is based on "**system thinking**":
  - Iterative co-design and co-optimisation loop, from system requirements down to the HW and spanning all the technology stack (or system layers).
  - Multidisciplinary team leveraging expertise in diverse areas.

Success lies in the ability to meet clients' computational needs effectively and seamlessly, minimising their effort to build on hardware.

Emerging alternatives?

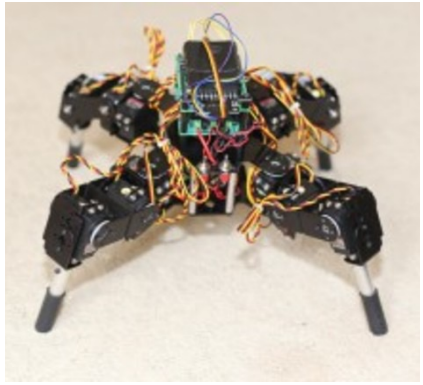
Semiconductor-based AI will be the main solution in the next 3-5 years, with a very good compromise (cost, efficiency, performance, etc.), for a very wide range of applications.

Bio-inspired systems

Cerebral organoids

3D structure grown from pluripotent stem cells in a lab, mimicking certain features of the human brain (e.g. cerebral cortex):

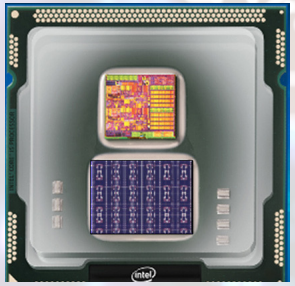
+ Programmable biological systems
+ Integrable with bio systems
+ More realistic cognitive processing
+ Integration with silicon
- Integrated in robots
  - Navigation
  - Object manipulation
- Not mature technology



Neuromorphic

Analog HW solutions mimicking biological neural networks (SNN), event-driven rather than continuous data processing:

+ Process information efficiently
+ Adaptable and flexible
+ Low power consumption
- New computing model
- Small dataset
- Non intensive tasks
- In-sensor solution
- Real time pattern recognition & decision making



Intel Loihi 2

Emerging alternatives? (2)

Photonics-based accelerators

Integrated photonics using optical interference to compute in parallel, quickly and efficiently:

+ Combine the precision of photonics with the practicality of CMOS manufacturing
+ Eliminate bottlenecks associated with electronic data transfer
+ High bandwidth, low latency, and minimal energy dissipation
− Integration of optical and electronic components challenges
− Scaling to complex NNs
− They will require new materials
• Neuromorphic + integrated photonics …
• Quantum photonics + integrated photonics …

Quantum technologies

Quantum computing has the potential to revolutionise AI leveraging the unique properties of quantum mechanics for learning and prediction.

− Futuristic approach for the edge
+ Increase parallelism
+ Accelerate and optimised training
+ Improve features extraction
+ Solve problems that are challenging or even intractable for classical computers

Thanks for the attention