Challenges and Opportunities to Enable

# Large Scale Computing via Heterogeneous Chiplets

Weifeng Zhang
Chief Architect & VP of Software, Lightelligence

In collaboration with
Zhuoping Yang, Shixin Ji, Xingzhen Chen, Jinming Zhuang,
Prof. Peipei Zhou, University of Pittsburgh
Dharmesh Jani, Meta

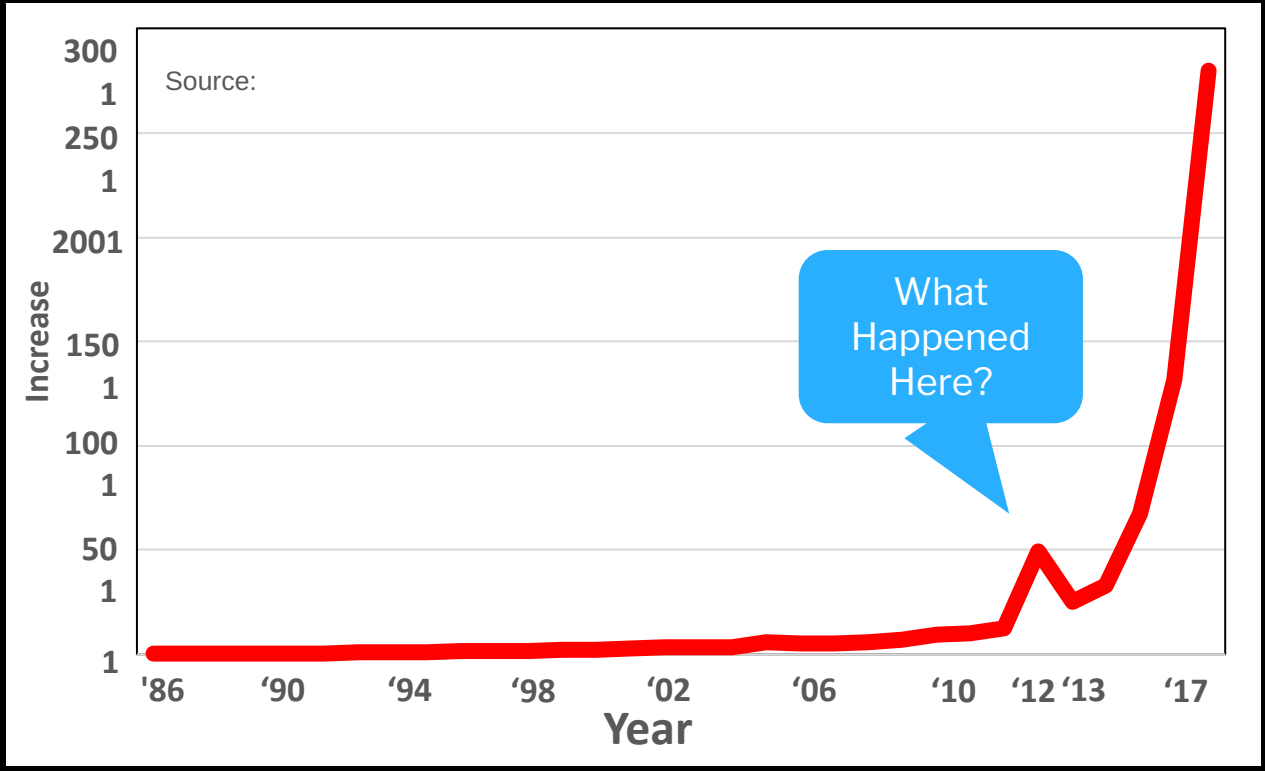
ASP-DAC 2024

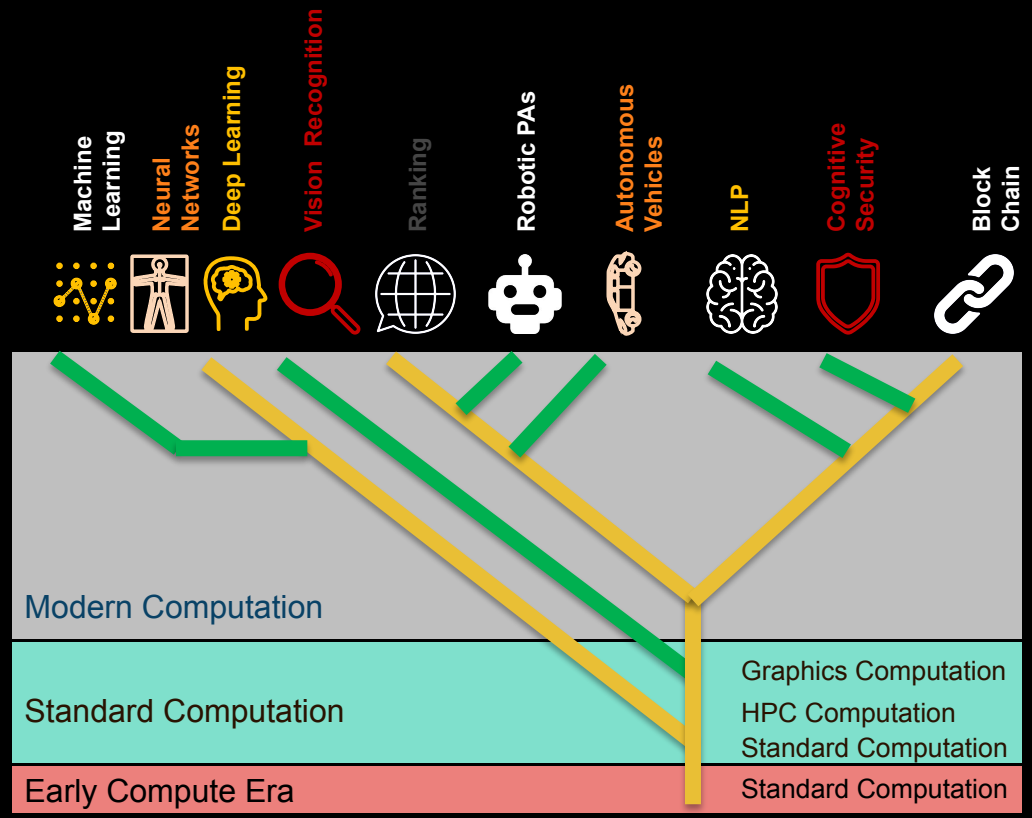
# Outline

- AI application characterization – true drive of demands
- Compute architecture evolution – taking on the demands
- Future of architecture – heterogeneous chiplets and design challenges

# Cambrian Explosion of Workloads

Growth of the Term “Deep Learning” in Research

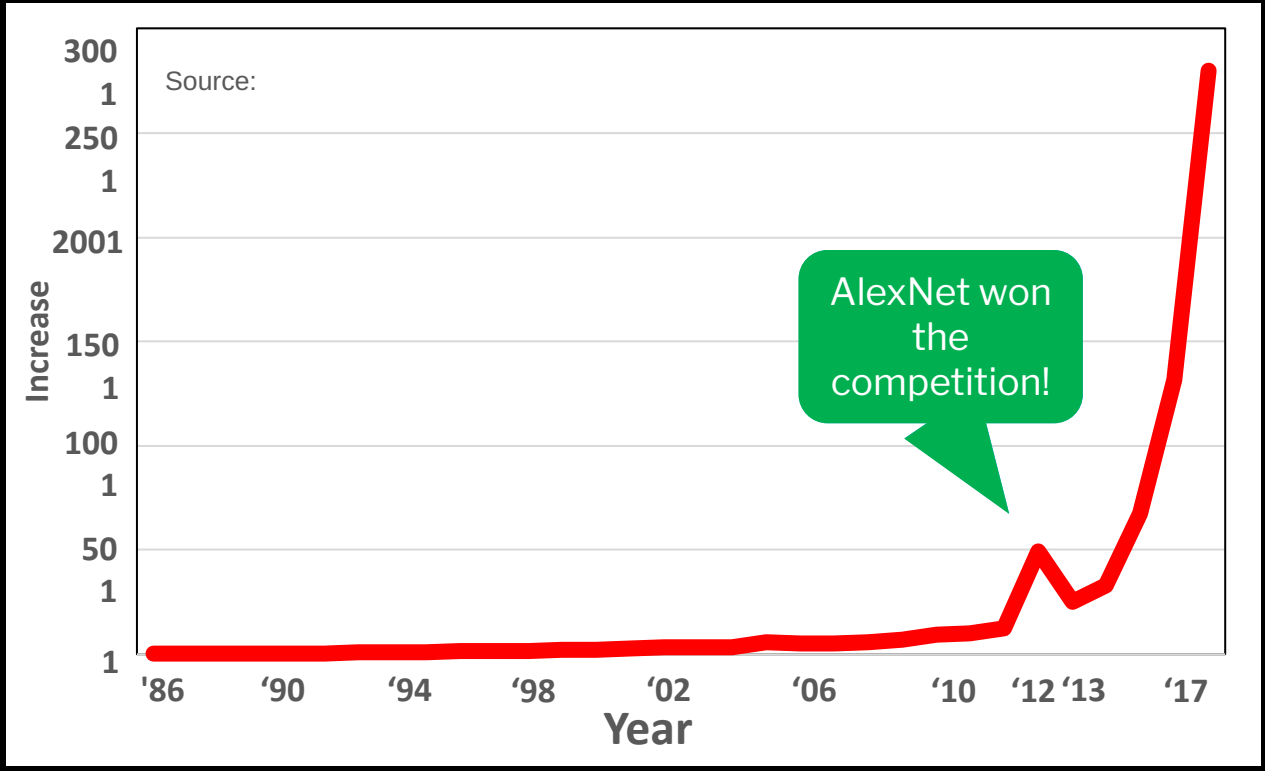


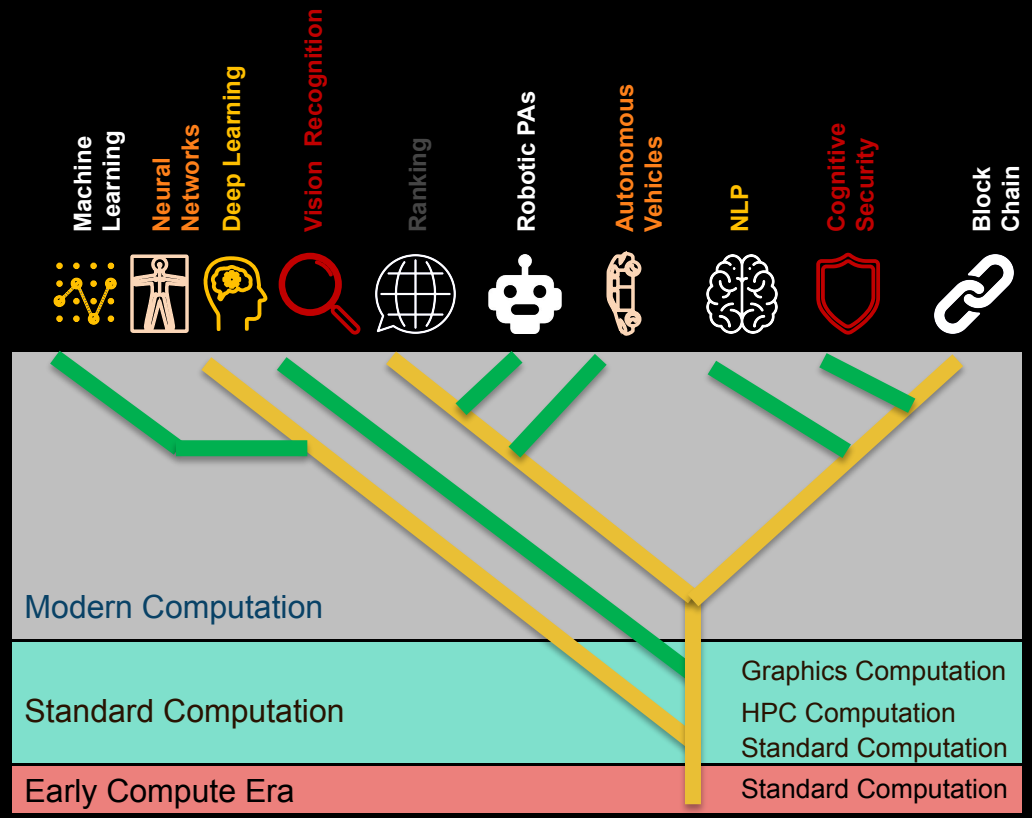




AI / Machine-learning and data-heavy workloads have exploded in 7 years and will diversify as new applications are discovered constantly…

# Cambrian Explosion of Workloads

Growth of the Term “Deep Learning” in Research

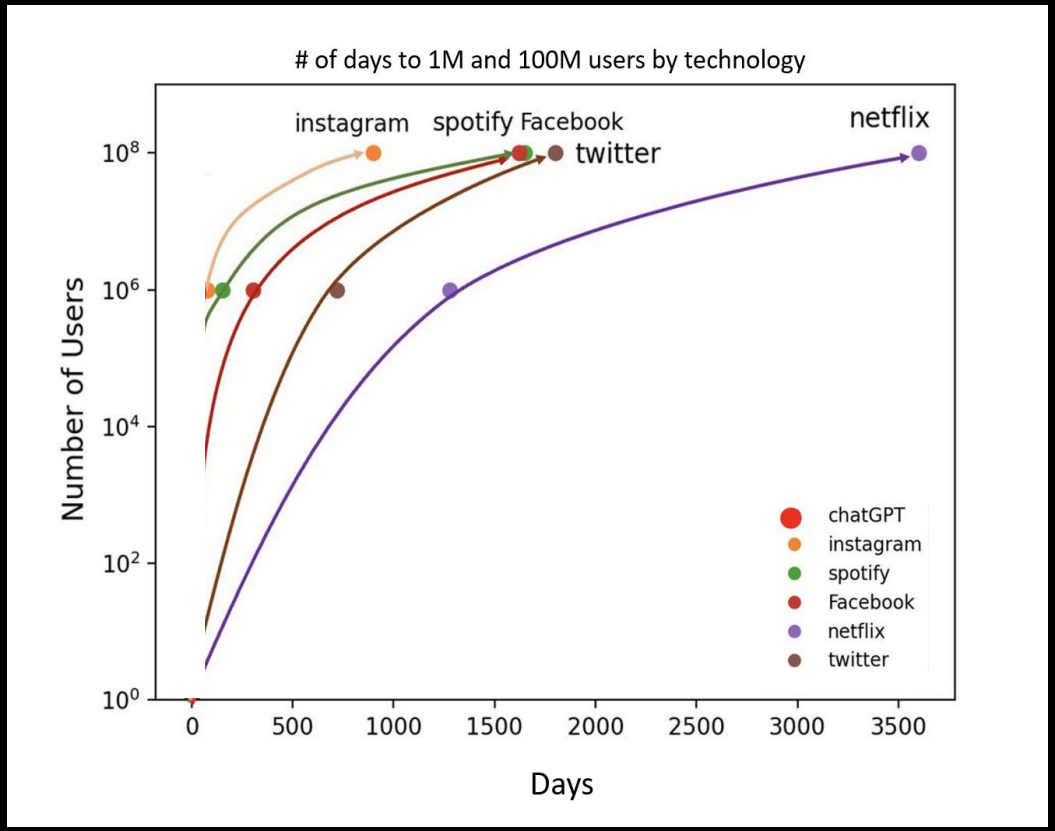


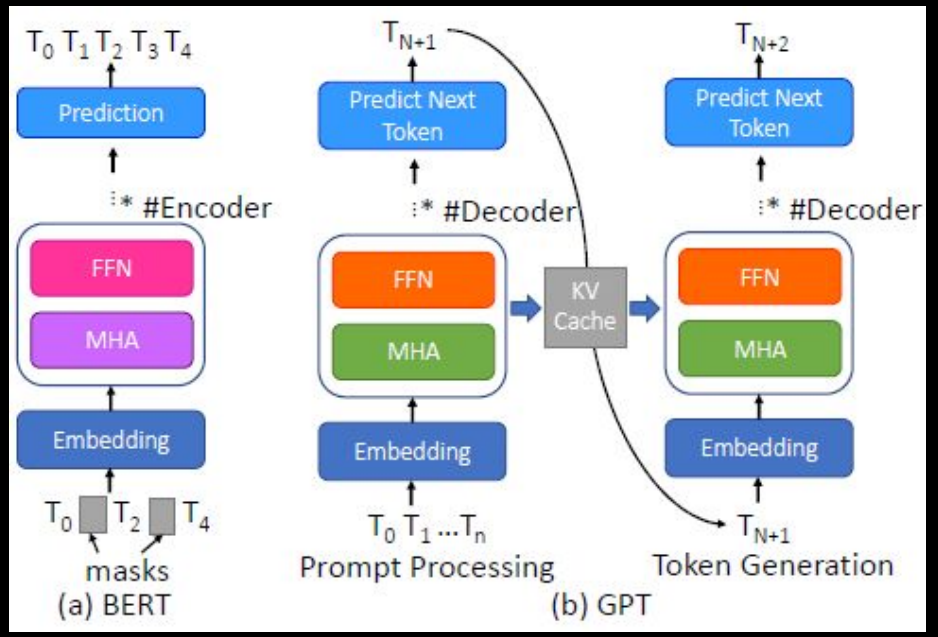




AI / Machine-learning and data-heavy workloads have exploded in 7 years and will diversify as new applications are discovered constantly…

# GEN AI as Emerging Workload

Success of ChatGPT launched on Nov 30th, 2022, redefined AI space

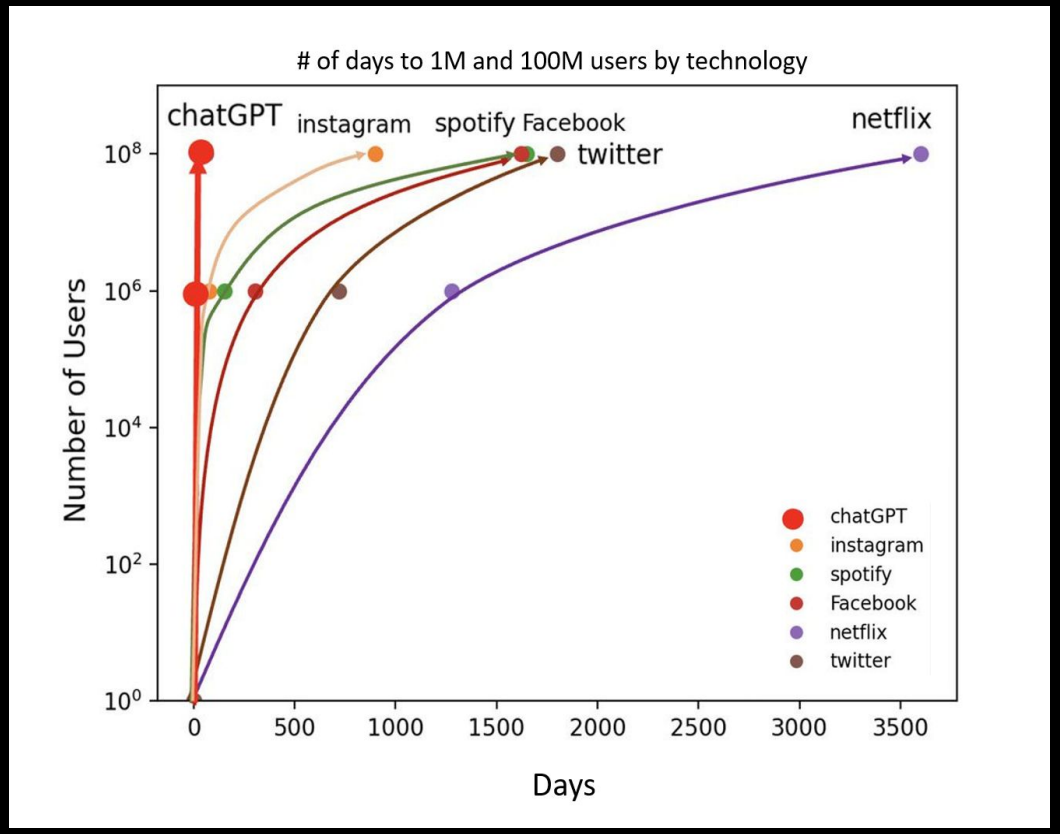


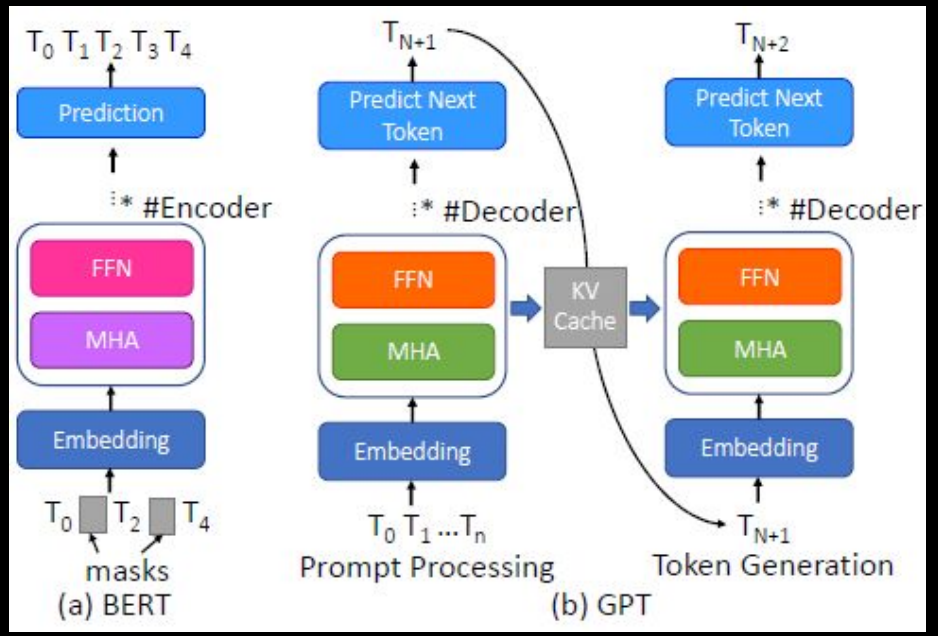




Different processes in BERT and GPT

# GEN AI as Emerging Workload

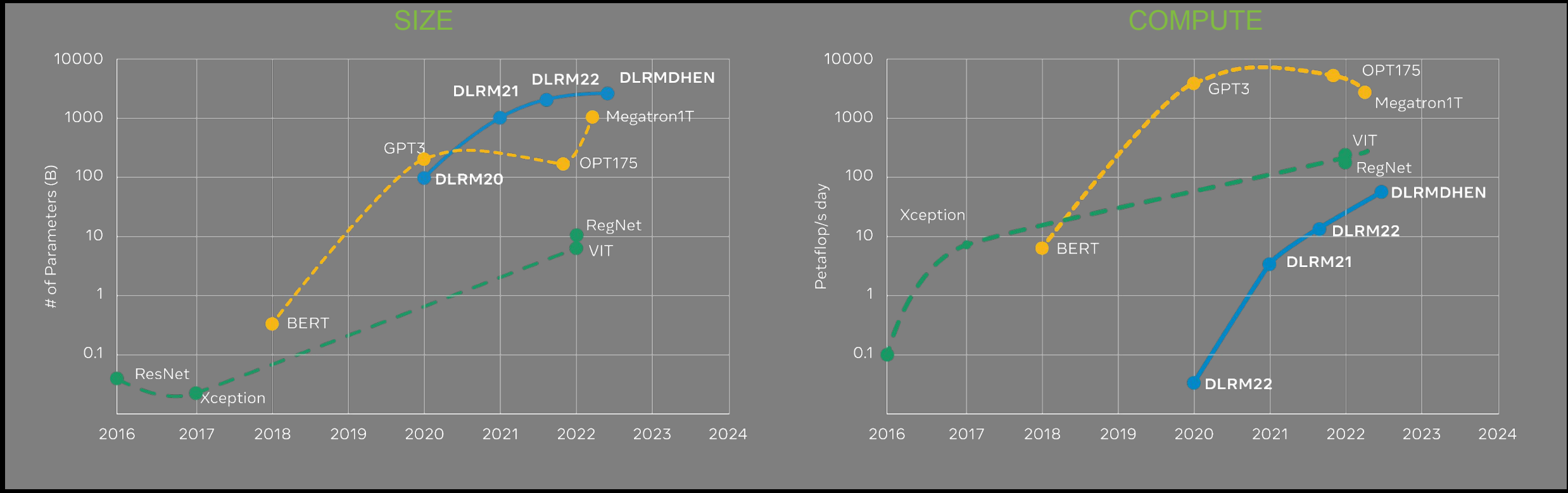
Success of ChatGPT launched on Nov 30th, 2022, redefined AI space



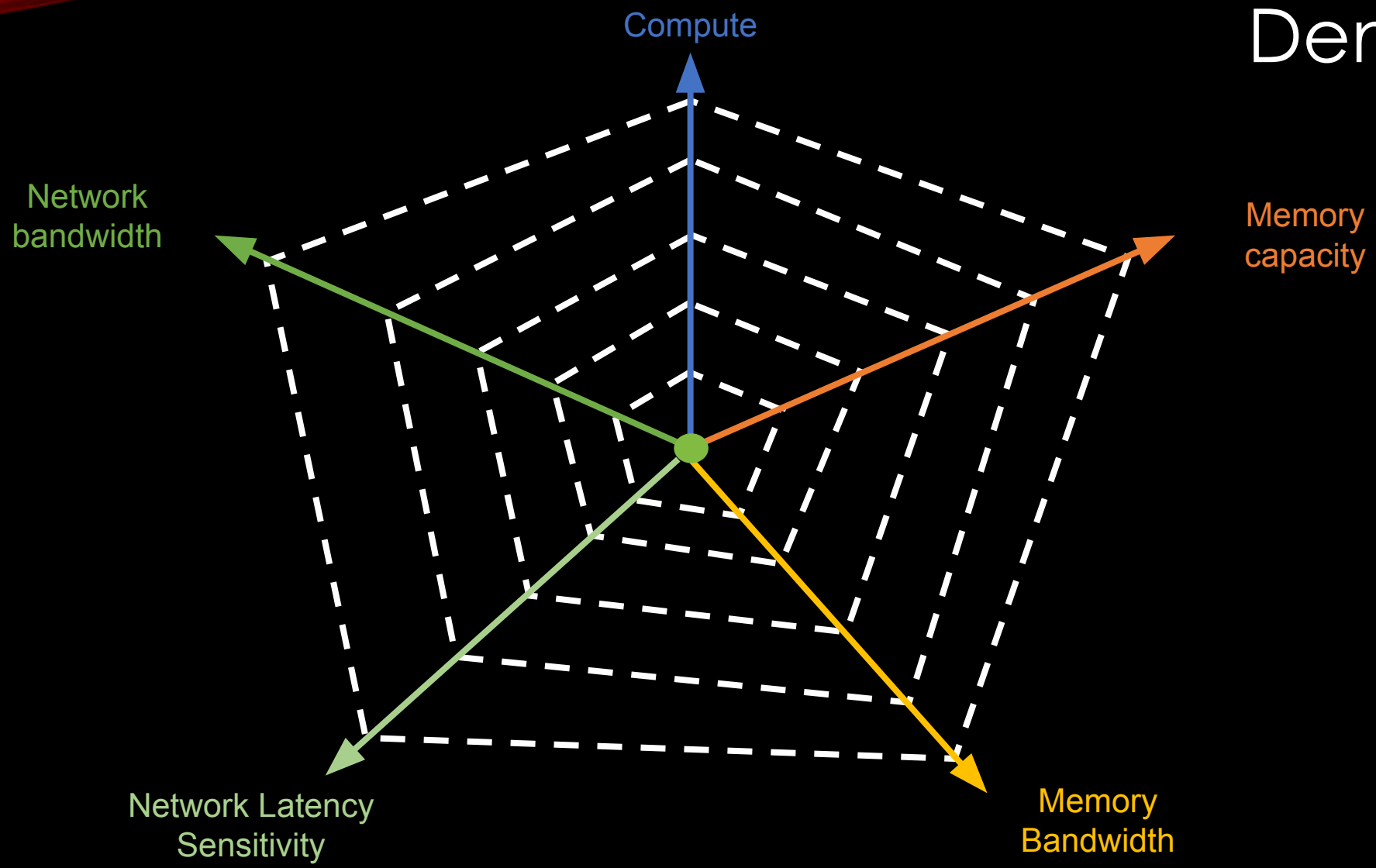

Different processes in BERT and GPT

# Deep Learning Workloads - Characteristics



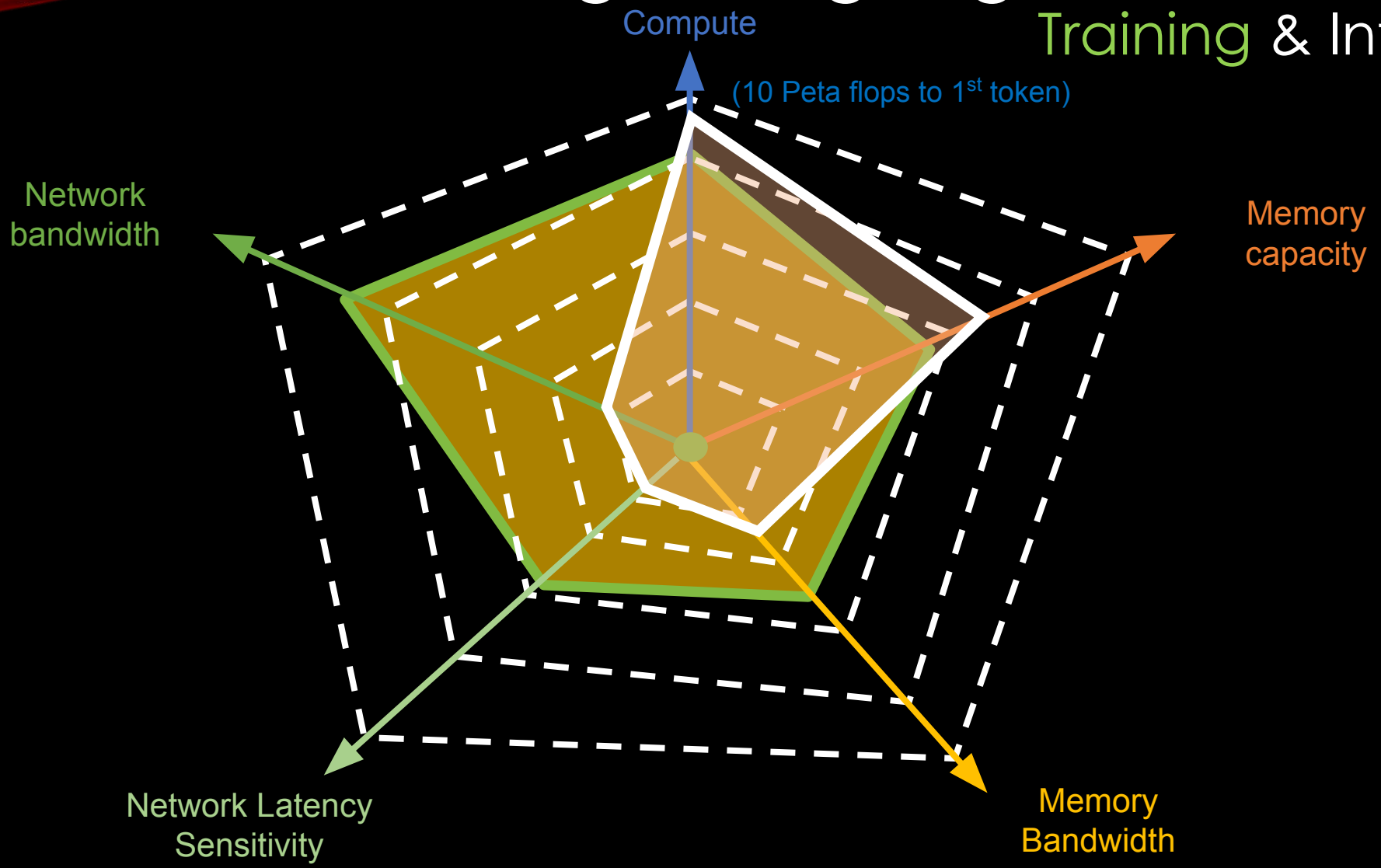

SOURCE: Meta at OCP Global Summit Oct 2022

# AI Workloads Impose Diverse Demands



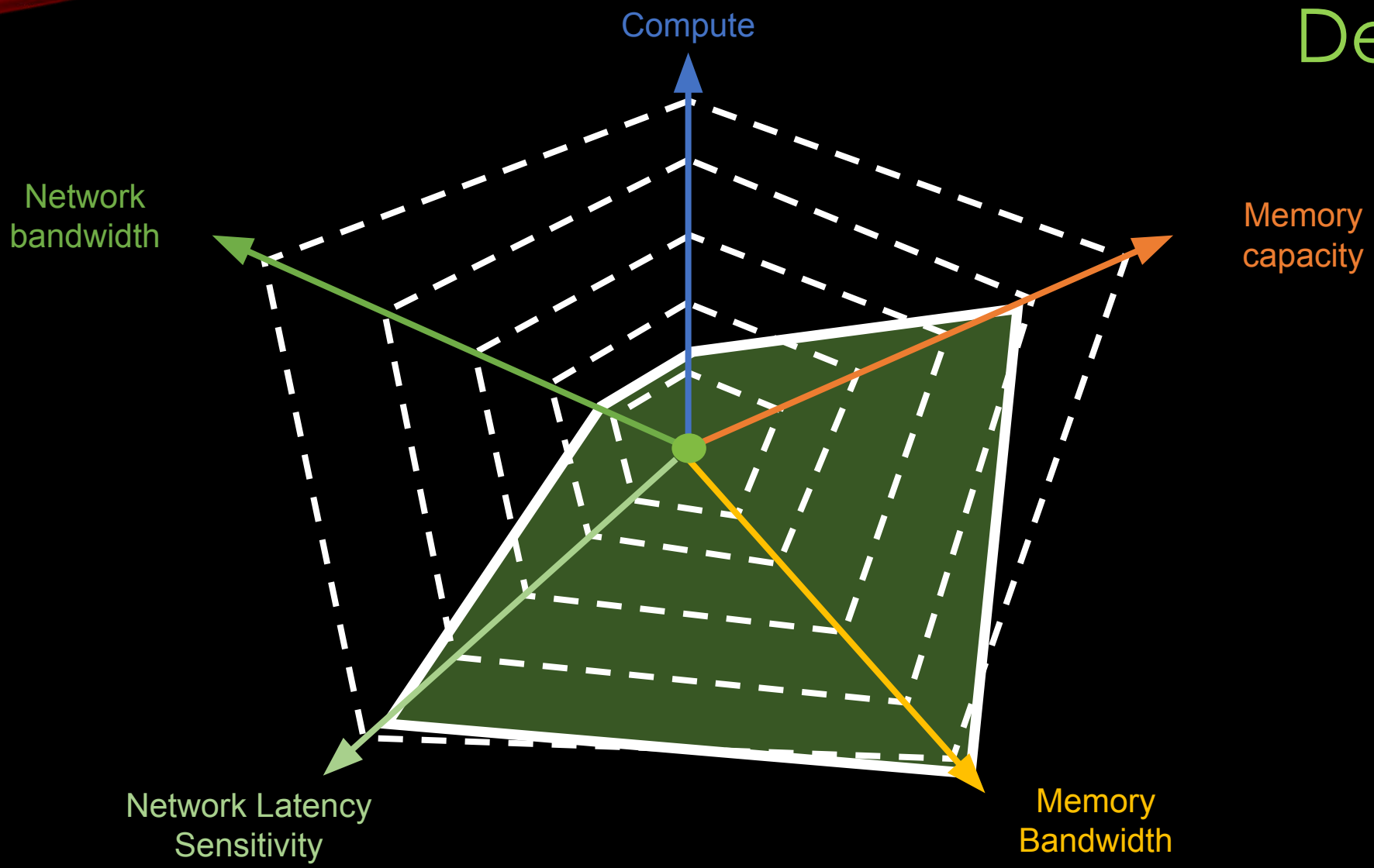

SOURCE: Meta at OCP Global Summit Oct 2022

# Large Language Models (LLMS)
## Training & Inference




SOURCE: Meta at OCP Global Summit Oct 2022

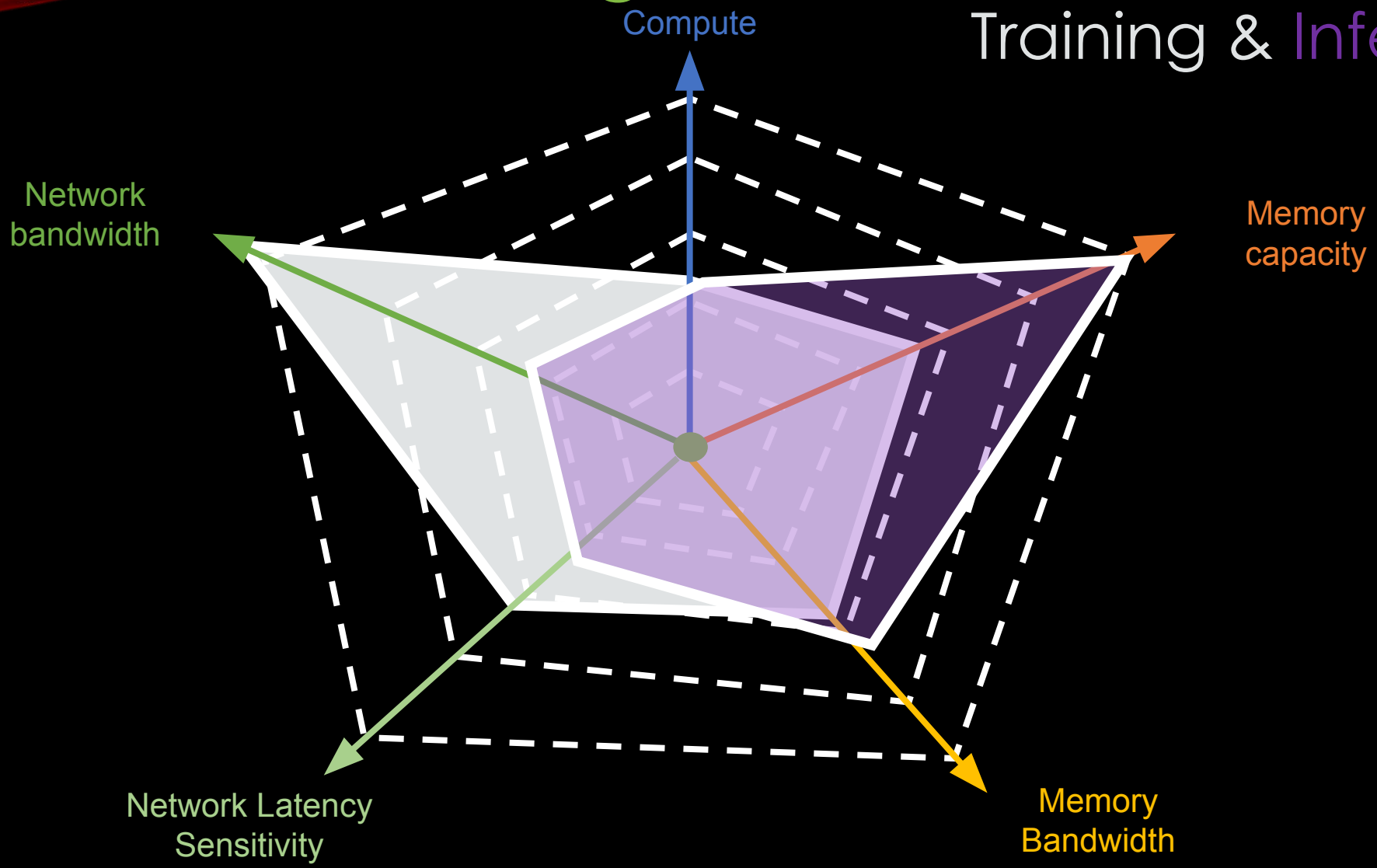
# LLMS Inference
## Decode




SOURCE: Meta at OCP Global Summit Oct 2022

# Ranking and Recommendation

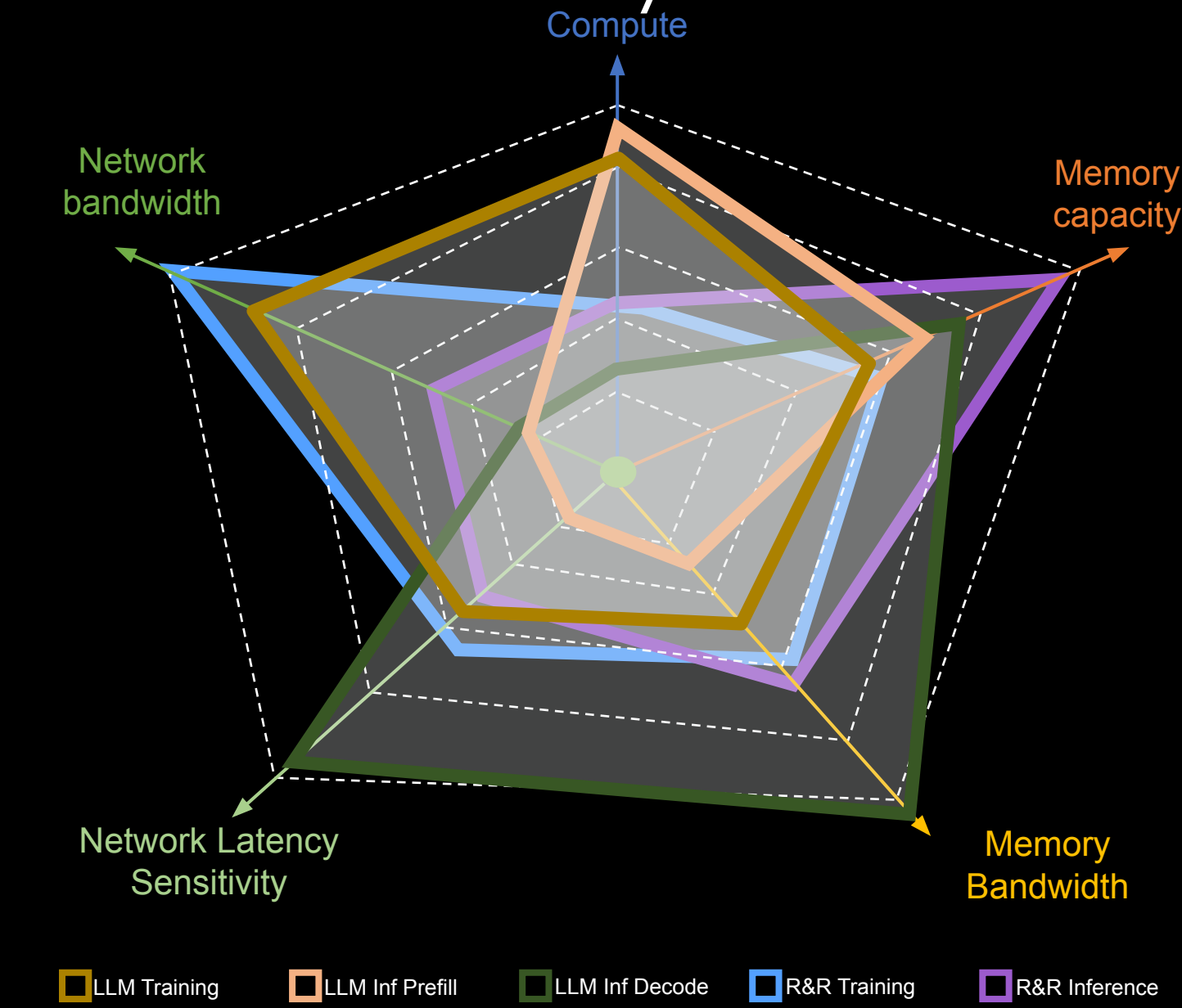
## Training & Inference




SOURCE: Meta at OCP Global Summit Oct 2022

# Workload Diversity Continues

- New models & parallelism techniques put unprecedented pressures on AI systems
- Difficult to serve all classes of models with a single system design point
- The next frontier of innovation is in heterogeneous computing and software/hardware co-design




SOURCE: Meta at OCP Global Summit Oct 2022

# Outline

- AI application characterization – true drive of demands
- Compute architecture evolution – taking on the demands
- Future of architecture – heterogeneous chiplets and design challenges

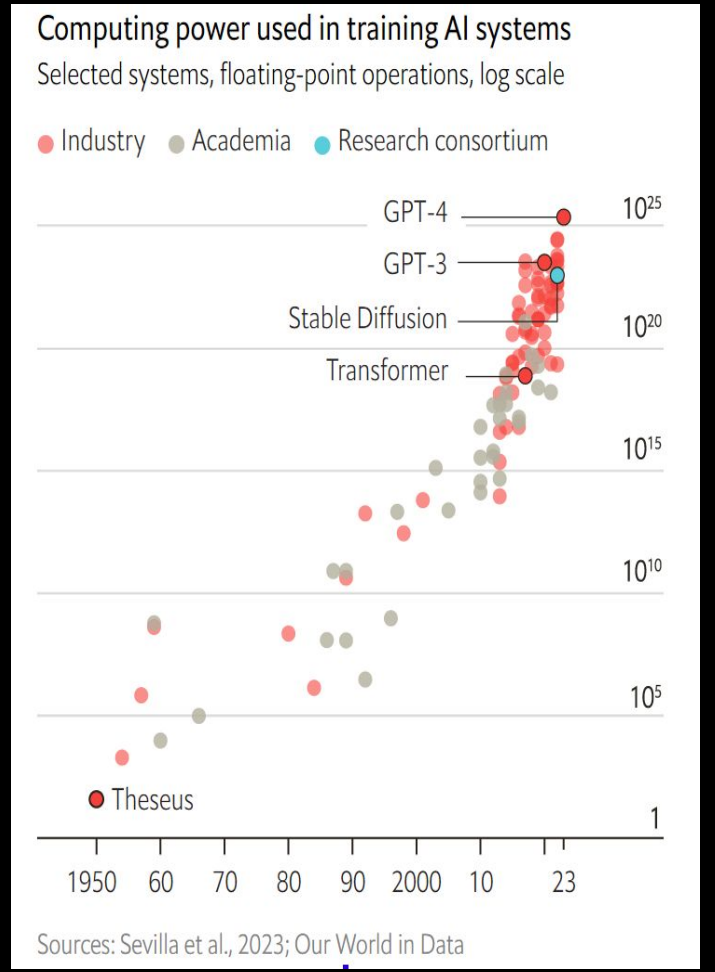
# Understanding LLM Models: Training and Inference Costs

| Model | Company | Parameters (p) (Billions) | Training Data (Token) (Billions) | Training Exa-FLOPS (6np) | Inference Tera-FLOPS (2*ni*p) | Memory Inference GB (p) | Memory Training GB (9p) |
|---|---|---|---|---|---|---|---|
| BERT | Google | 0.3 | 137 | 246.6 | 0.6144 | 0.3 | 2.7 |
| GPT-J | EleutherAI | 6 | 402 | 14472 | 12.288 | 6 | 54 |
| GPT-3 | OpenAI | 175 | 300 | 315000 | 358.4 | 175 | 1575 |
| **LLama-65B** | **Meta AI** | **65** | **1400** | **546000** | **133.12** | **65** | **585** |
| PaLM2 | Google | 340 | 3600 | 7344000 | 696.32 | 340 | 3060 |
| GPT-4 | OpenAI | 1000 | 20000 | 120000000 | 2048 | 1000 | 9000 |

$$\text{Training FLOPs} = O(6*n*p)$$
$$\text{Memory Training} = O(9*p)$$

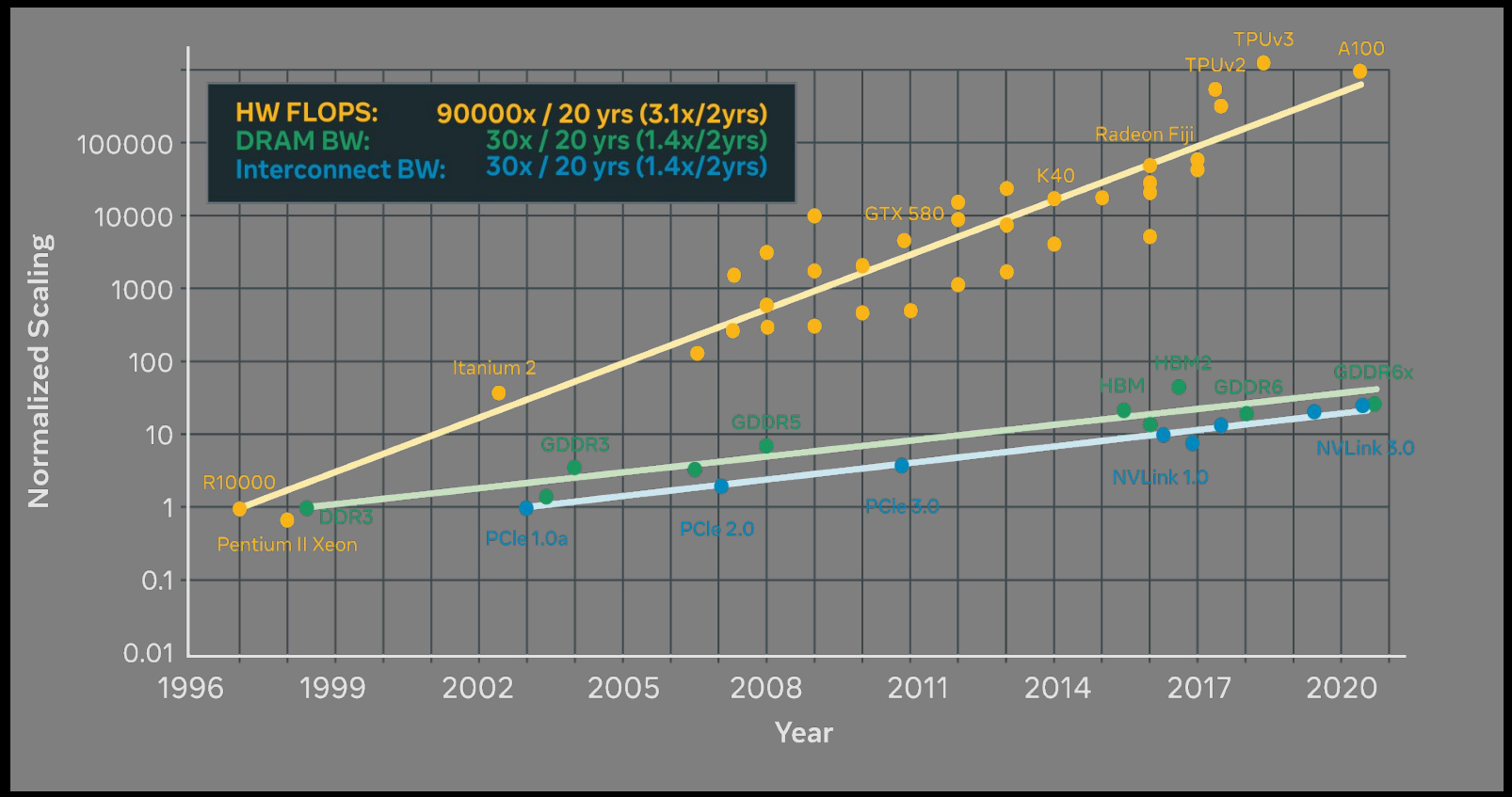
$$\text{Inference FLOPs} = O(2*n_i*p)$$
$$\text{Memory Inference} = O(p)$$

p = Model parameters
n = Size of tokens used for training
$n_i$ = Size of tokens used for inference



# Compute Architecture for AI Workloads

- Hardware accelerators (GPUs, ASICs, FPGAs) for throughput and energy efficiency
  - GPU: massive parallelism for large batches of training data
  - ASIC: better customization for low latency in real-time scenarios
  - FPGA: trade-off of programmability, performance, and quick prototyping
  - Trend of heterogeneous integration




Scaling of peak HW FLOPS and memory/interconnect bandwidth

# Scaling up Computation with Chiplets

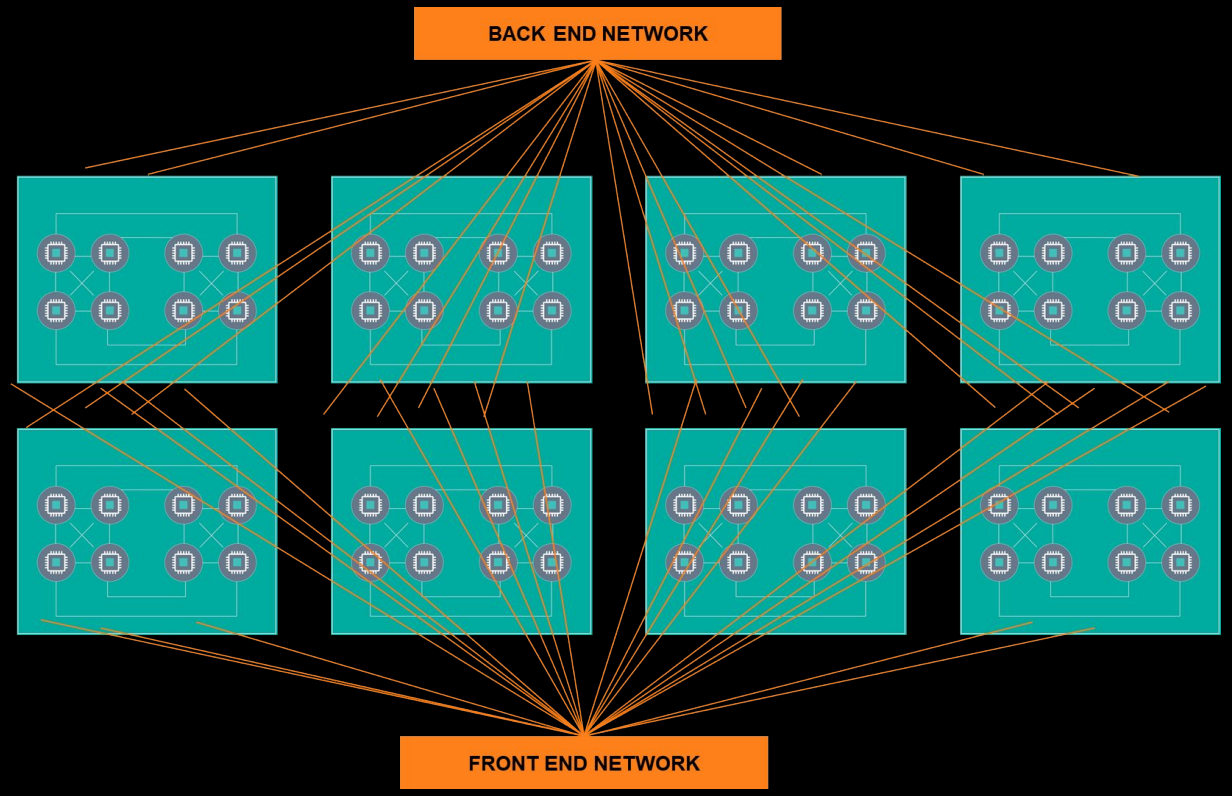
- Integration with chiplets
  - Overcome lithographic reticle limits and yields (design complexity, manufacturing cost, and integration density)
  - Heterogeneous integration of IPs with mature processes
  - Flexible selection of chiplets for different customer requirements
  - Reuse/cost and time-to-market

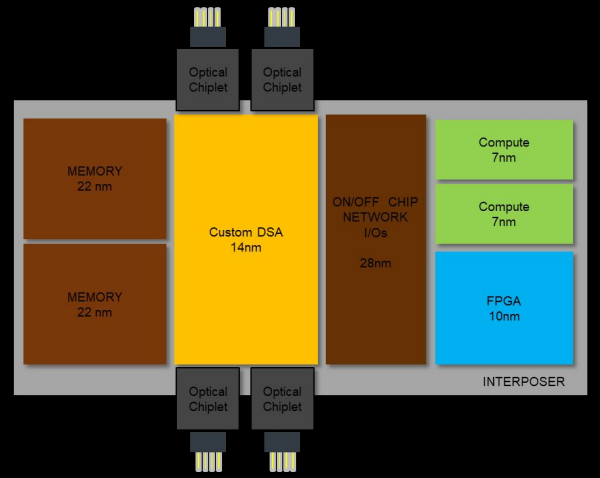
TABLE I: Comparisons between chiplet and PCB, monolithic ASIC [10].

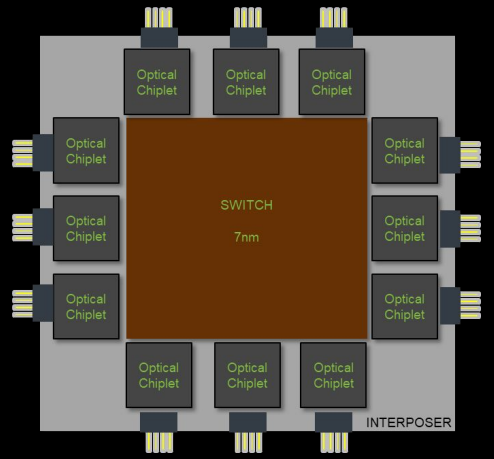
| Integration Technology | Design Cycle | Cost/$ | Integration | Energy cost | Performance |
|---|---|---|---|---|---|
| Monolithic ASIC | >1 year | >1,000,000 | +++ | + | +++ |
| Chiplet | months | 1,000-1,000,000 | ++ | ++ | ++ |
| PCB | weeks | 100-10,000 | + | +++ | + |

# Scaling up Interconnection with Chiplets

## Training with DSA Clusters




Low power high radix opto electrical switch

# Chiplets for AI Systems: Challenges (1)

- Chiplet interface
  - Interconnect protocol and standardization
  - Routing algorithms in SiP: passive vs. active interposers
  - Pre-silicon hardware simulation

TABLE II: Comparisons between different chiplet interfaces (data accessed in 2023/11).

| Protocol | Institution | Typical Energy Efficiency (pJ/bit) | Maximum Speed (Gbps/wire) | Fault Tolerance Mechanism |
|---|---|---|---|---|
| USR [21] | \ | <0.6 [21] | >20 [21] | N/A |
| AIB [22], [23] | Intel | 0.85 (Gen1) | 2 (Gen1) [22], 6.4 (Gen2) [23] | N/A |
| BoW [24], [25] | ODSA | <0.25-1.0 [25] | 32 [24] | N/A |
| HBM [26] | JESDC | \ | 6.4 | ECC |
| LINPINCON [27] | TSMC | 0.424 | 2.8 | N/A |
| UCIe [28] | UCIe Union | 0.25-1.25 [28] | 32 GT/s [28] | CRC + Retransmission |
| AAC [29], [30] | China Chiplet Industry Alliance | 2.5 [29] | 128 [30] | CRC + BER + Retransmission |

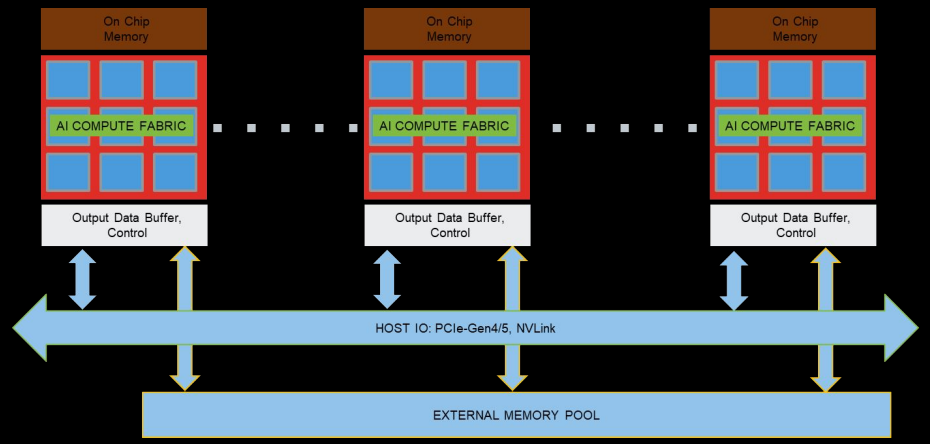
# Chiplets for AI Systems: Challenges (2)

- Packaging related issues:
  - Testing – still open area to describe chiplet testing pins, coverage, standard interface
  - Thermal – online management & offline optimization avoiding dark silicon issues
  - Co-design – need standards to describe chiplet electrical properties and fast simulation for holistic design flow

# Chiplets for AI Systems: Challenges (3)

- Security related issues:
  - More vulnerable to security threats
    - During chiplet interaction or integration
    - Complex architecture and mixed trust environment
  - Potential threats:
    - Side channel, fault injection, hardware Trojan
  - Potential protection methods:
    - Trusted execution environments (TEEs)
    - Root of Trust via active interposer
    - Chiplet-based hardware security module

# Chiplets for AI Systems: Challenges (4)




Training on DSAs

- Orchestration and portability of AI workloads
  - Conflicts of AI runtimes
  - Different code development environment
- Unified programming infrastructure for chiplet:
  - SYCL: one API – a whole software stack
  - MLIR: ScaleHLS to enable hierarchies of design and larger design space optimization
  - HeteroCL decoupling algorithm specifications
- Software tools:
  - Task partitioning and mapping (e.g., H2H, CHARM)
  - Design space exploration

# Acknowledgement


- Authors from University of Pittsburgh were also sponsored by National Science Foundation: NSF #2213701, #2217003, #2324864, #2328972