Infrastructure AI technology
Accelerated by Advanced Memory & Flash Technology

Jung H. Yoon, Ph.D.

Distinguished Engineer & CTO

Supply Chain Engineering
IBM Infrastructure

Outline

1. Infrastructure technology driving forces
2. Si technology & Advanced Packaging
3. Memory Technology
4. Flash Storage Technology
5. Power subsystem & Interconnect
6. Summary

Infrastructure Technology – Driving Forces

Performance

CPU/GPU Si scaling, Heterogeneous integration

Low latency - AI training & inferencing

Bandwidth – Memory, Interconnect, Optical Transceivers, PCIe Gen 5/6

Density – Memory, Storage Capacity &

I/O Density

Quality & System Resilience

Design for quality

Manufacturability & Reliability

Sub Tier Quality

FW Quality

System RAS

Sustainability

Power Consumption TCO

Carbon footprint

Security

Supply Chain Security

Product Security

Ransomware detection

Supply Chain Resilience

Geo-political factors

Component design

Cost

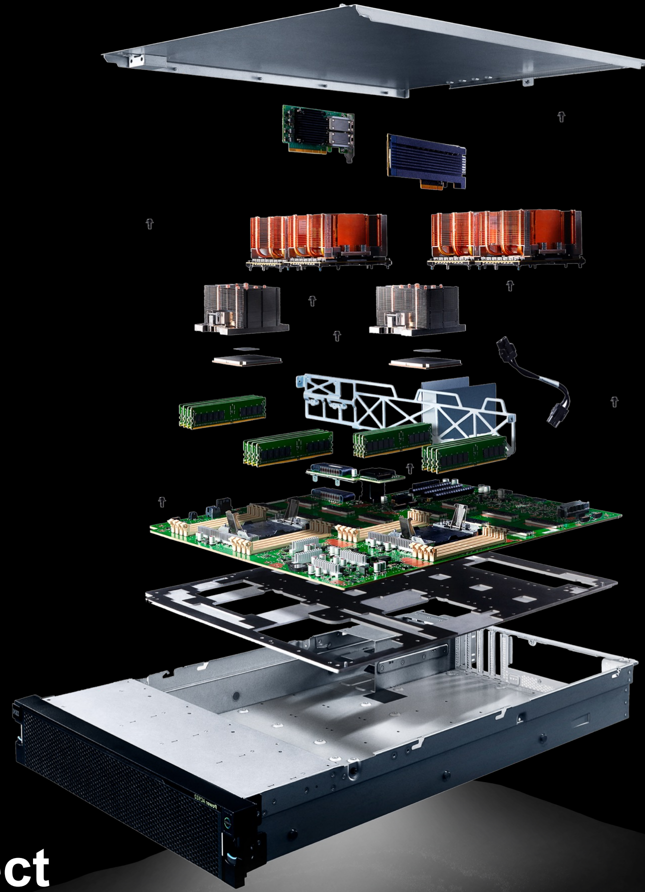
Semiconductor scaling

CXL disaggregated infrastructure

the Future of Memory and Storage

Infrastructure – Building Blocks

1. Processor & GPU
2. Memory
3. Flash / SSD
4. HDD & Tape
5. Power Subsystem
6. Interconnect



13. Switch
12. Power Components
11. Thermal/Cooling
10. Mechanical
9. ECAT
8. PCB
7. Fiber Optics

AI placing huge demands to Infrastructure

100x
more model parameters

AI training and inferencing will require highly intensive processing of simultaneous computations

10x
growth in newly-generated AI data

The iterative supply chain of synthetic, annotated and generative data that must be stored, secured, managed

7x
faster security threat lifecycles

Bad actors are using AI to identify new data and AI
model vulnerabilities and to speed iterative attacks

7x
greater computational throughput

Elaborate data-intensive models will exponentially scale the rate of storage, memory and processors transactions

50x
performance degradation from distributed data

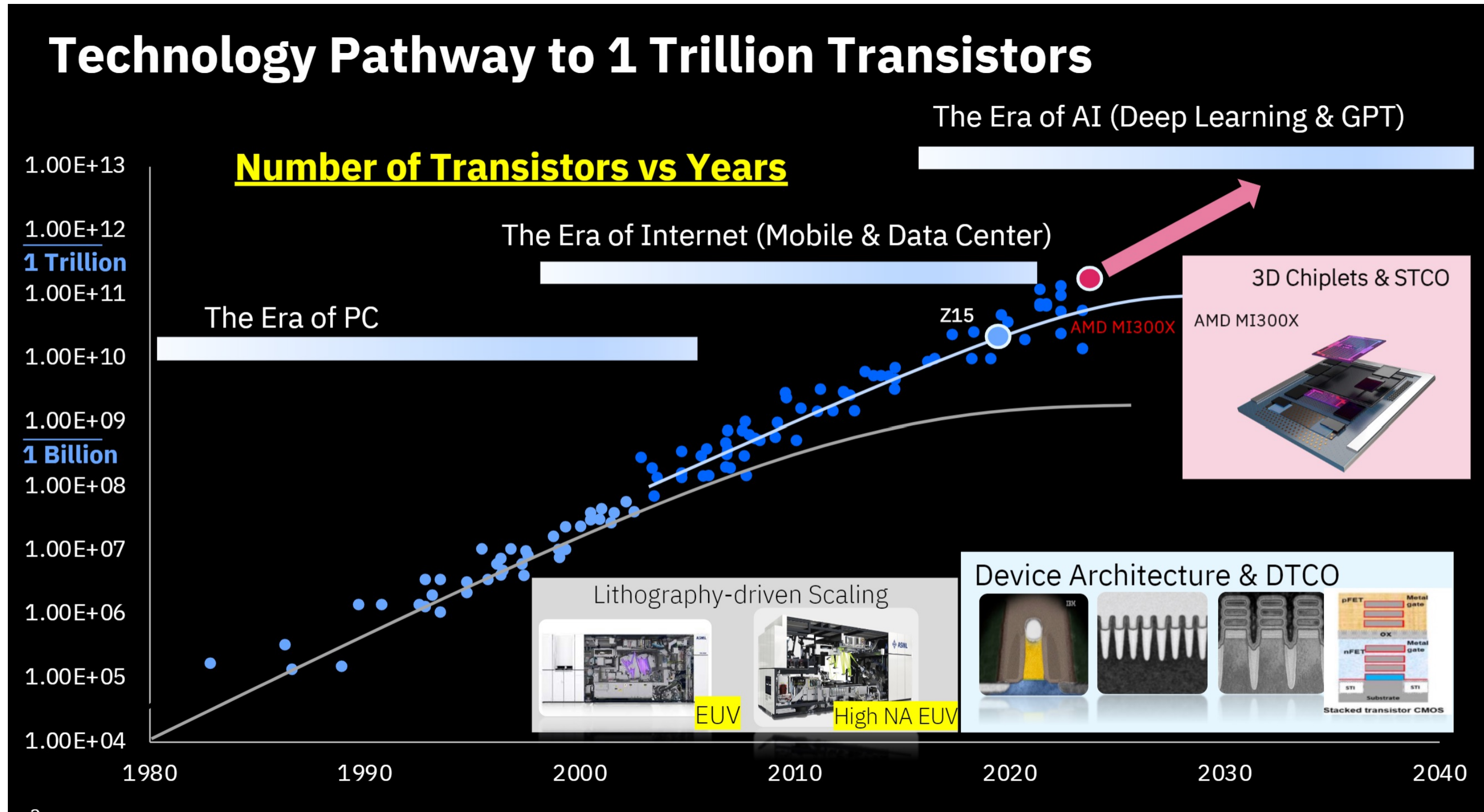
Driving the need to dynamically distribute AI models to where the data lives to achieve near-zero latency and scalable responsiveness

2x
more energy consumption

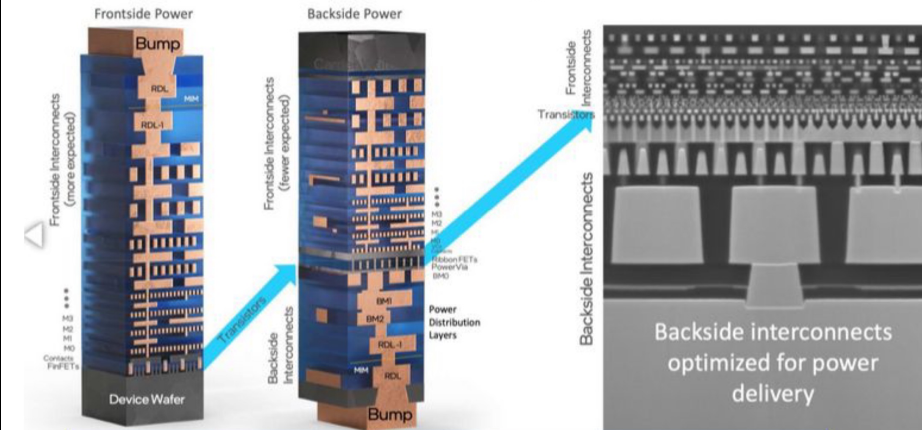
As AI models rapidly consume more and more data, iteratively trained longer and longer and require faster and faster inferencing cycles

AI fueling Si Scaling & Advanced Packaging

- Vertical transistor architectures & back side wafer interconnects
- 3D Chiplets, High Bandwidth Memory TSV stacked packaging
- AI is driving all aspects of Compute: >40 Peta FLOP performance, 48GB HMB4 cube with 16High TSV stack



D. Chibambarrao, IEEE EDTM 2024



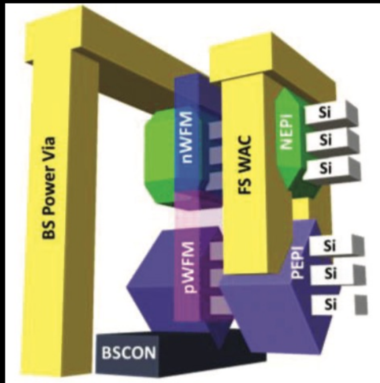
W. Hafez, VLSI 2023

the Future of Memory and Storage

AI stimulating Power & Thermal Solutions

the Future of Memory and Storage

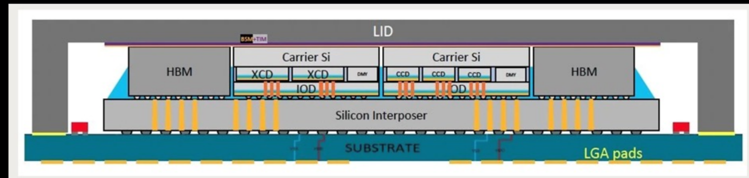
Chip



Radosavljevic et al, IEDM, 2023

Package

AMD MI300



https://www.amd.com/en/products/accelerators/instinct/mi300.html

https://spectrum.ieee.org/amd-mi300

System

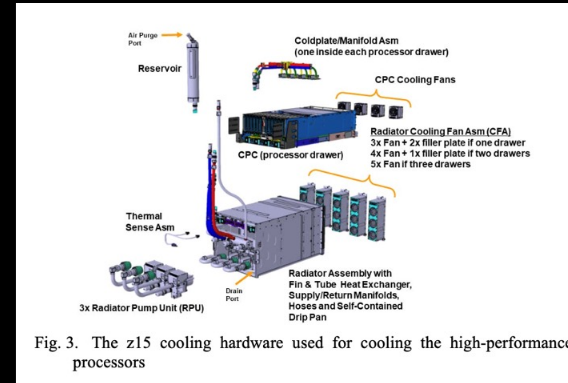
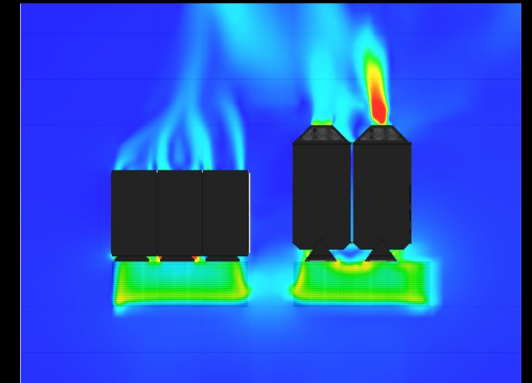


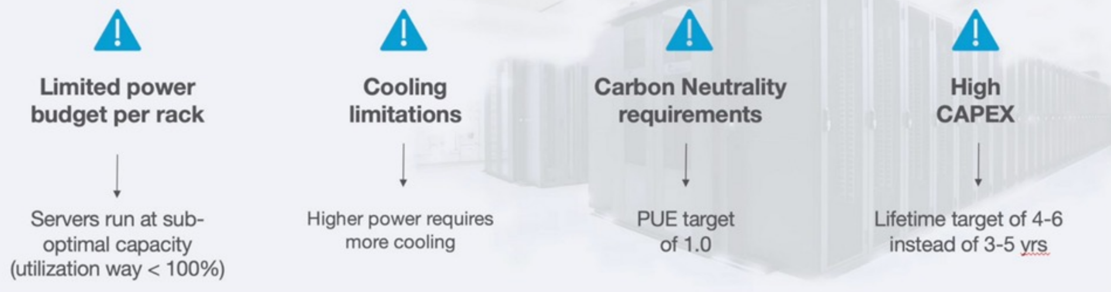
Fig. 3. The z15 cooling hardware used for cooling the high-performance processors

Demetriou et al, iTherm, 2021

Data Centers



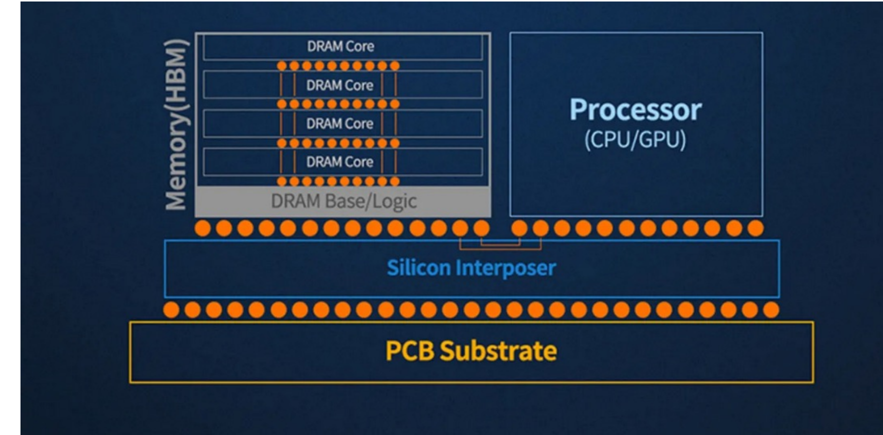
Demetriou & Hu: IBM z15 Data Center Digital Twin Models



Source: protean Tecs

- **AI intensive workloads:** more Si processing devices > more interconnects > more memory
- **Stacked and Packed:** MORE POWER & HEAT generated
- AI requires **System level solution** – solving components independently cannot produce optimal design point
- **Performance/power** driving new levels of Si technology, Chip mechanical integrity & reliability => Thermal Solutions are fundamental

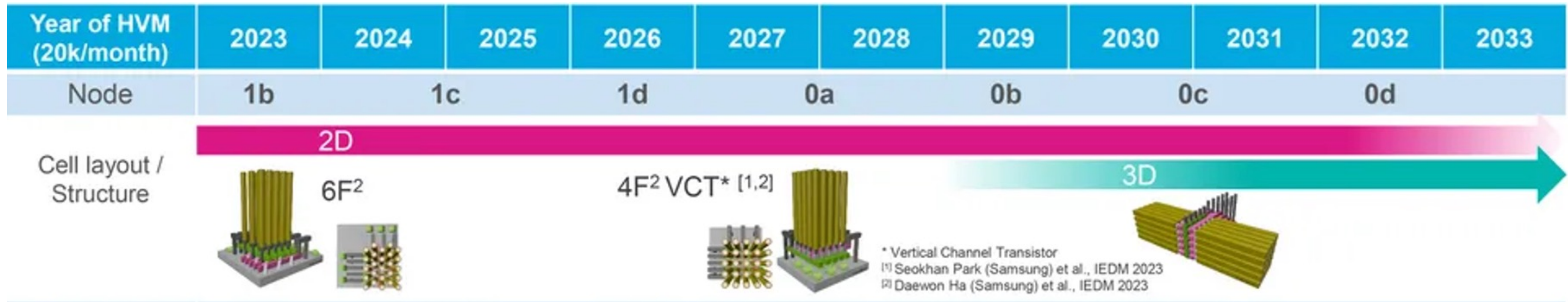
Memory Technology – AI directions



- **Training and Inferencing**
  - GenAI driving highest memory capacity & Bandwidth
  - Growth of model parameters, Need to store Large Language Model
- **HBM (High Bandwidth Memory)**
  - **TSV (Thru Si Via) stacking** allows layers of Si die to be stacked on top of each other
    - Allows large number of memory chips to be packed into smaller volumetric space – reduce the distance that needs to travel between the Processor/GPU to memory (e.g., HBM4 max capacity 48GB, Max B/W 1.65TB/s, 16H TSV)
  - **Lower power consumption** – reduce the amount of power needed to transfer data between memory to processor
    - HBM can reduce heat generated by memory – enables overall system performance & reliability

Memory Technology Scaling – AI viewpoint

- DRAM transistor scaling – 6F2 planar > 4F2 vertical transistor, 3D-DRAM enabling i) device performance scaling, ii) bit cost scaling, iii) transistor power scaling, iv) density scaling
- Wafer bonding technologies – to optimize array efficiencies, cell performance
- Memory intensive AI workloads, power scaling, thermal engineering, new reliability mechanisms



Source: TEL Investor Relations, Feb 2024

Flash Storage Technology – AI Infrastructure

1. **Data volume**
    - AI workloads require massive amounts of data for ingestion, training, validation and inference
    - Gen AI LLM driving increased flash memory capacity & Bandwidth
    - 1-2 Tb QLC enabling high capacity data lakes
        - Data lakes more read intensive with lower write requirements

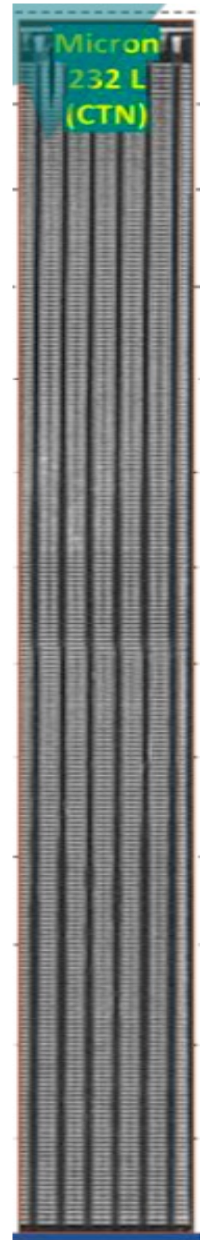
2. **Data Velocity**
    - The rate at which data is generated and processed increasing tremendously for AI.
    - Real time analytics & decision making - critical for finance, healthcare, cloud/edge.
        - High speed data ingestion, processing and retrieval critical to enable real-time AI applications

3. **High Performance**
    - AI workloads require Flash enabled high-performance storage - meet computational demands of training complex models and processing large data sets
    - Low latency access to data, and high throughput to support intensive AI workloads and maximize compute infrastructure investments. Random access latency performance – key for training & inferencing

4. **Data Efficiency**
    - Critical for optimal performance, energy conservation for accelerated training and inference, scalability and reliability.
    - Power and space efficient storage important for overall infrastructure efficiency & TCO

Flash Technology Scaling – AI implications

- **Growing #of layers** enabling 3D NAND manufacturers to drive Density & $/GB scaling
  - 3D-NAND scaling 3xx layer MP in 2025, 1000 layer tgt'd by 2030
- **Z-height reduction** critical for continued 3D-NAND scaling – Channel etch control, cleaning, stress management, die thinning & stacked package quality
- **3D-NAND architecture** – block size, # of planes, page size
  - Cell current challenges resulting in larger block size vs 3D-NAND generation – complexity in wear leveling & overall flash management
- **Wafer fab process & W2W bonding technology**
  - Wafer bonding allows flexibility in optimizing cell performance, array efficiency, and interface speed
- **Key performance parameters** – tRead, tProg, tBERS key in driving performance
- **Reliability** – endurance, data retention, read disturb vs 3D-NAND scaling



Techinsights
FMS2023, J. Choe

FMS
the Future of Memory and Storage

FMS the Future of Memory and Storage

Power Subsystem

- AI infrastructure requires **higher power capacities** > 3000 Watt ratings to meet AI centric power requirements
- **Power Density** – wide band gap GaN technology best solution for high density & efficiency
- **Energy efficiency** – more efficient PSU & DC/DC converter designs to minimize energy waste and TCO
- **Reliability** – IBM SystemZ requires eight 9 resilience
- **Redundancy & fault tolerance** – ensure continuous operation and data processing via redundant PSU configurations and DC/DC converter designs
- **Smart power management** – real time monitoring/telemetry, power usage tracking

Interconnect

- **High B/W:** Data intensive AI workloads requiring massive computing power and storage capacity
  - 224 Gbps-PAM4 development pushing the laws of physics for both semiconductors and interconnect
    - Pushing limits of design & manufacturing
- Co-development of multiple disciplines– Hardware, architecture, connectivity, mechanical integrity, signal integrity

Summary

1. **AI demands to Infrastructure**
   - Model parameters, data size, computational thruput, near zero latency, scalability, security, Power consumption

2. **Gen AI driving forces - key technical directions**
   - **Si technology** – sub 5nm scaling, 3D chiplets, Heterogeneous integration
   - **Memory** – High BW memory, DRAM transistor scaling
   - **Flash** - Tb QLC, data lake application
   - **Power subsystem** – high wattage, energy efficiency, fault tolerance
   - **Interconnect** – high BW 224 Gbps PAM4

3. **Power Consumption & Sustainability**
   - Increased power trends driven by Gen AI – technology innovations needed across infrastructure
     - device scaling, power subsystem, thermal/cooling technology