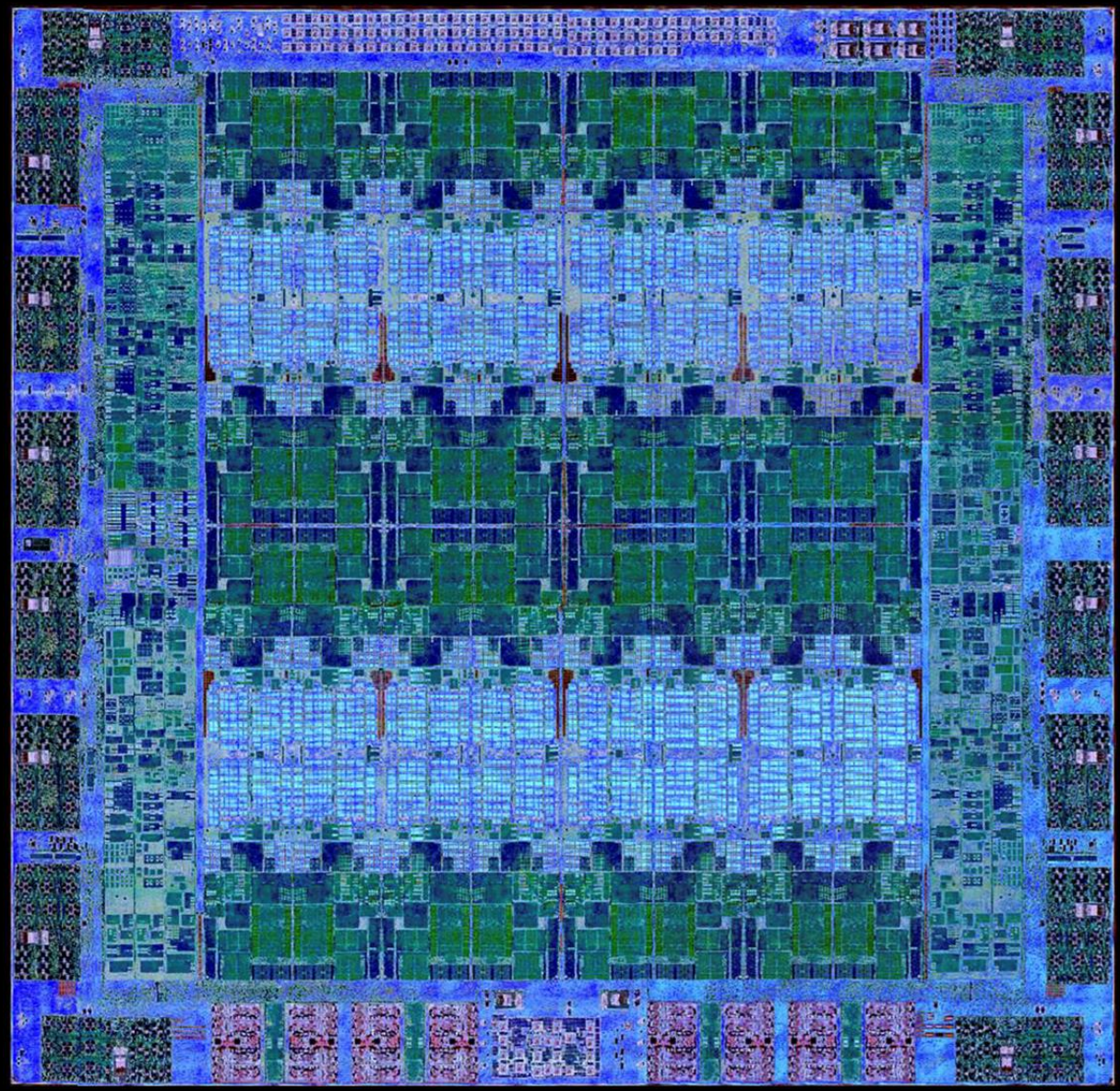
Power Processor Roadmap

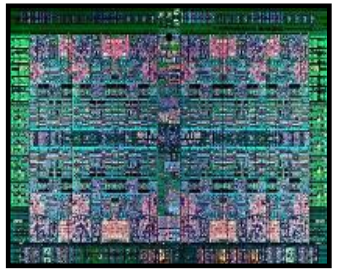
William Starke
IBM Distinguished Engineer
POWER Processor Chief Architect



IBM Power Processor Roadmap

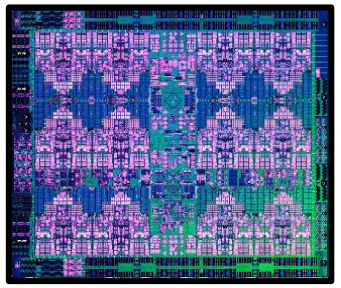
Continuous Platform Innovation and Leadership

POWER8



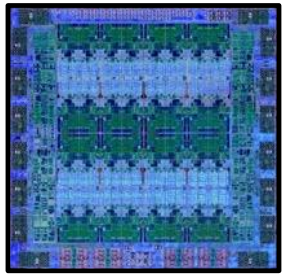
Powerful SMT8 Core
Enterprise Scaling
Big Data Optimized
Agnostic Memory

POWER9



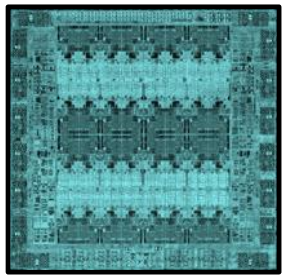
Modular Core Design
Accelerator Attach
(NVlink, OpenCAPI)
Data Plane Bandwidth
DDR & CDIMM Memory

Power10



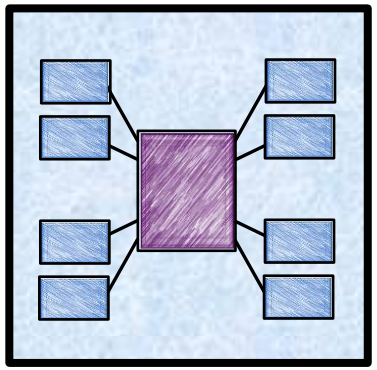
Core / Thread Strength
Socket Performance
In-Core AI Acceleration
Efficiency / Sustainability
HW Accelerated Security

Power11



Core / Thread Strength
Socket Performance
Enterprise Scaling
Energy Optimization
Improved up-time
Full Breadth AI

Power Future

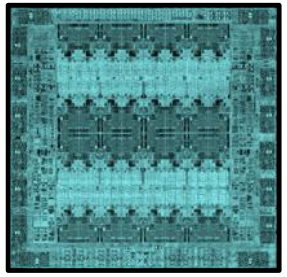


(Under development)

Information is forward looking and plans are subject to change based upon IBM's discretion

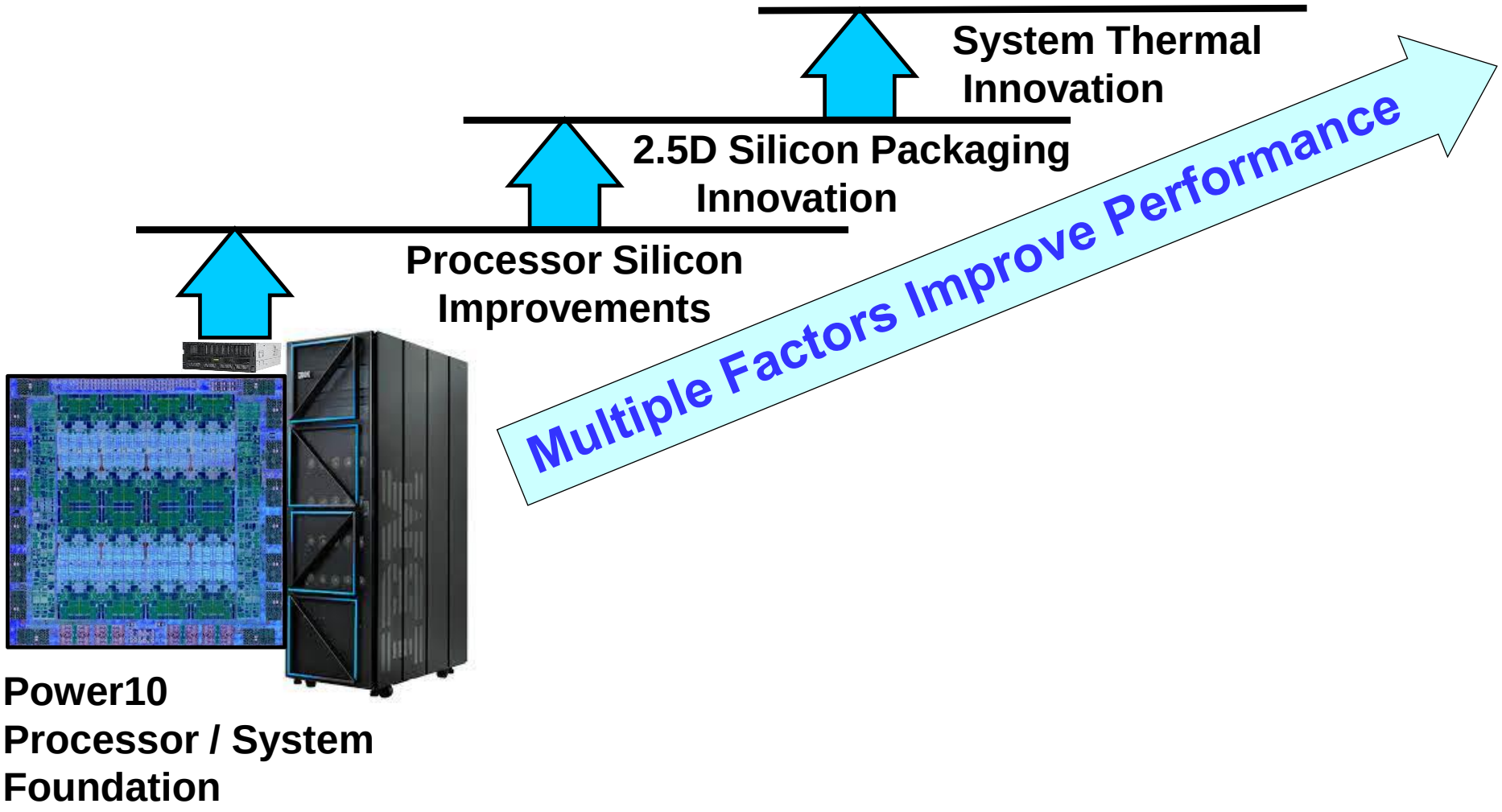
Power11: Full Stack Innovation and Cross-Optimization

Processor Architecture
Socket-level Packaging
Semiconductor Technology



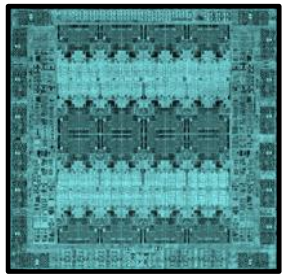
Improved Thread, Core, Capacity
ISC Silicon Layer: Energy Optimization
Samsung Foundries Enhanced 7nm

Processor Silicon Optimization: Improved Performance & Energy Mgmt, Socket TDP Innovation
Socket Packaging Innovation: 2.5D silicon stacking → Optimized Energy Delivery
System Thermal Innovation: Improved Cooling Enabling Higher Throughput



Power11: Full Stack Innovation and Cross-Optimization

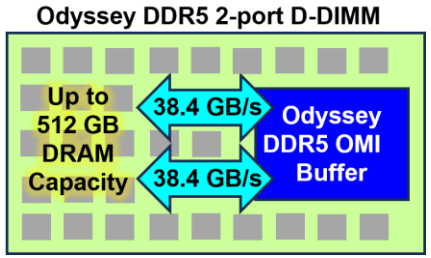
Processor Architecture
Socket-level Packaging
Semiconductor Technology



Improved Thread, Core, Capacity
ISC Silicon Layer: Energy Optimization
Samsung Foundries Enhanced 7nm

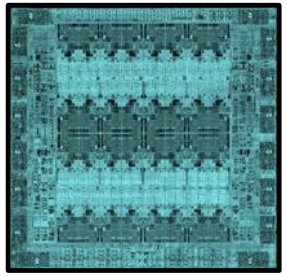
Power11: Full Stack Innovation and Cross-Optimization

Memory Architecture
Energy / Thermal Infrastructure



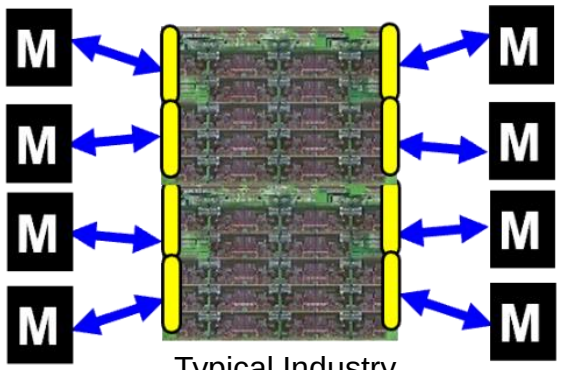
Agnostic, 3x Pipes, 2x Capacity
Advanced Cooling Technology

Processor Architecture
Socket-level Packaging
Semiconductor Technology



Improved Thread, Core, Capacity
ISC Silicon Layer: Energy Optimization
Samsung Foundries Enhanced 7nm

IBM Power and Z Buffered Server Memory



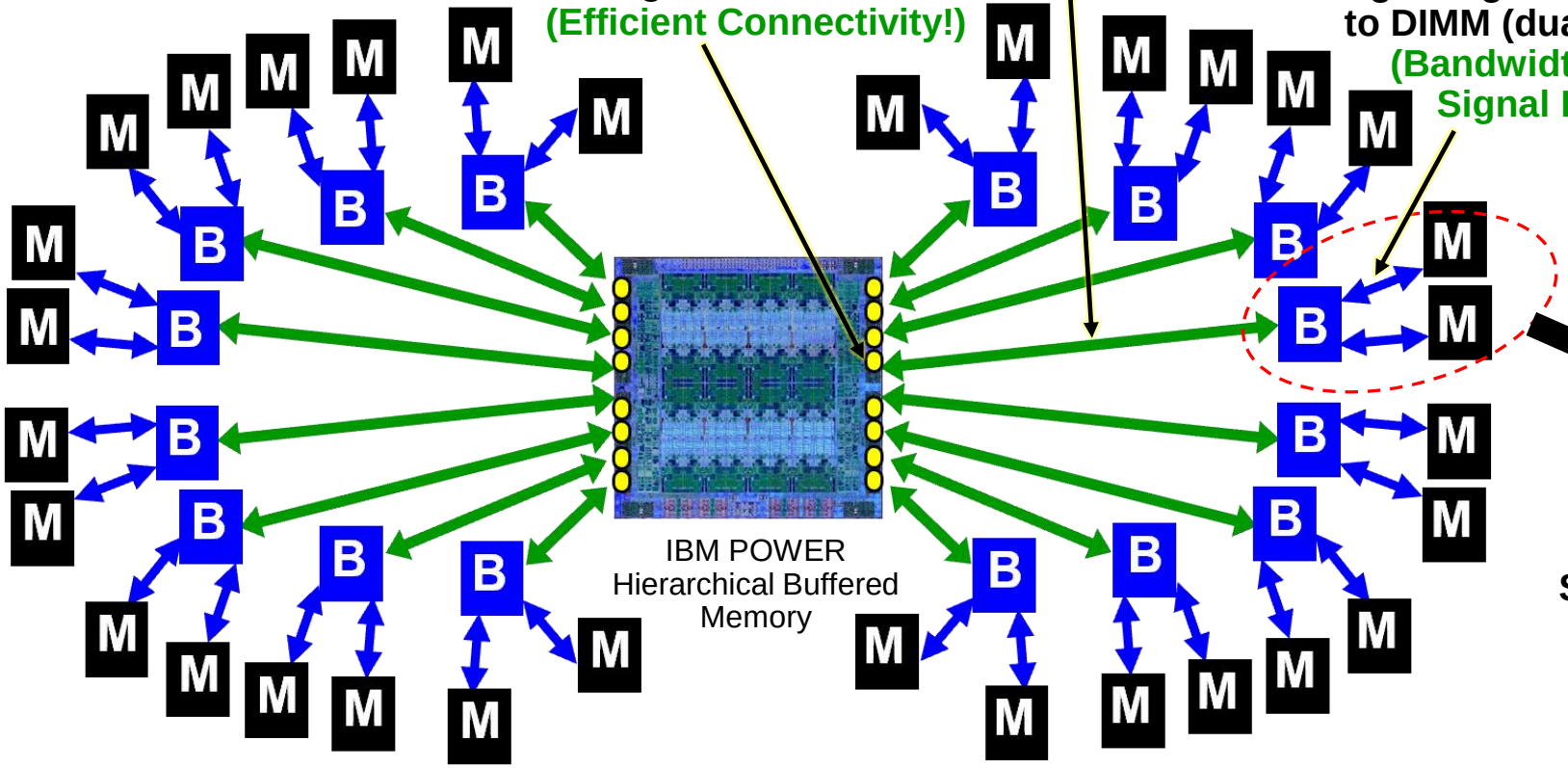
Typical Industry
Direct Attach Memory

Differential, ultra-high bandwidth signaling running packetized, replayable, steerable, agnostic, high-level OMI semantics from processor to laminate to motherboard to DIMM Connector to DIMM to buffer chip
(Excellent bandwidth and Signal Integrity!)

4x Bandwidth / Processor!
4x Capacity / Processor!
Superior Reliability!
9x Processor Area Value!
Technology Agnostic!!!

OMI port: 9x Bandwidth per processor chip edge area vs DDR
(Efficient Connectivity!)

DDR DRAM signal-ended signaling running contained to DIMM (dual port fan-out)
(Bandwidth, Capacity, Signal Integrity!)



IBM POWER
Hierarchical Buffered
Memory

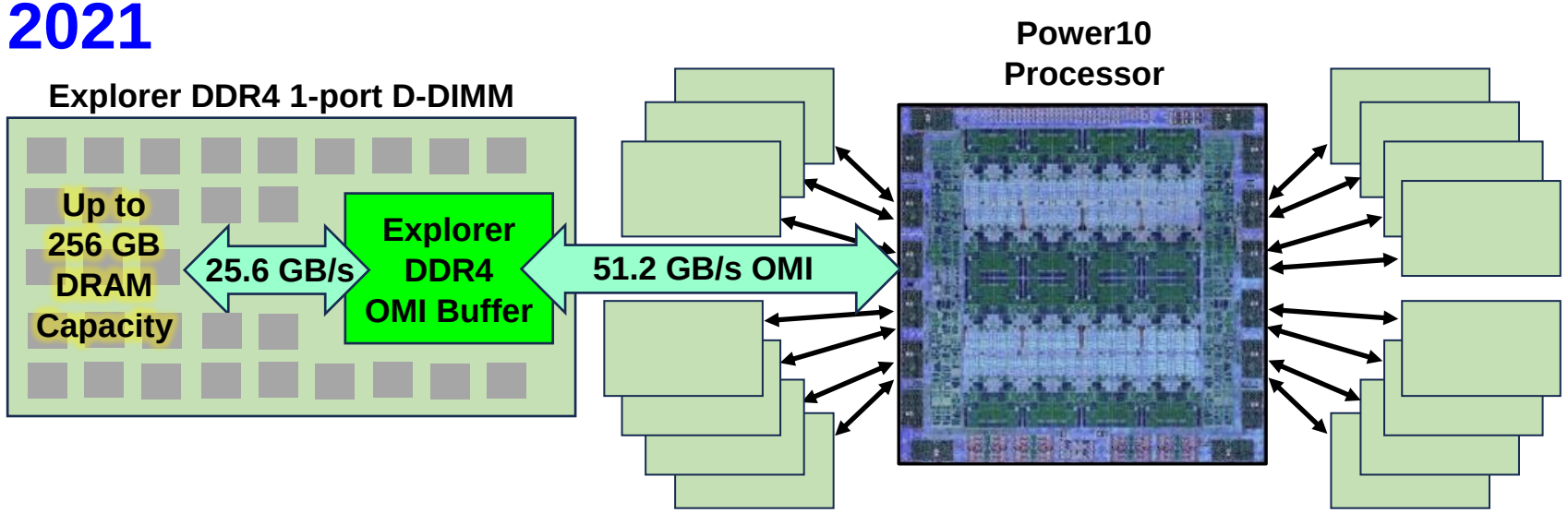
Spare DRAMs
(Reliability!)

DIMM Connector
(Protocol provides Resiliency!)

OMI D-DIMM

Memory Technology: OMI DDR5 Performance, Scaling, Economics

2021

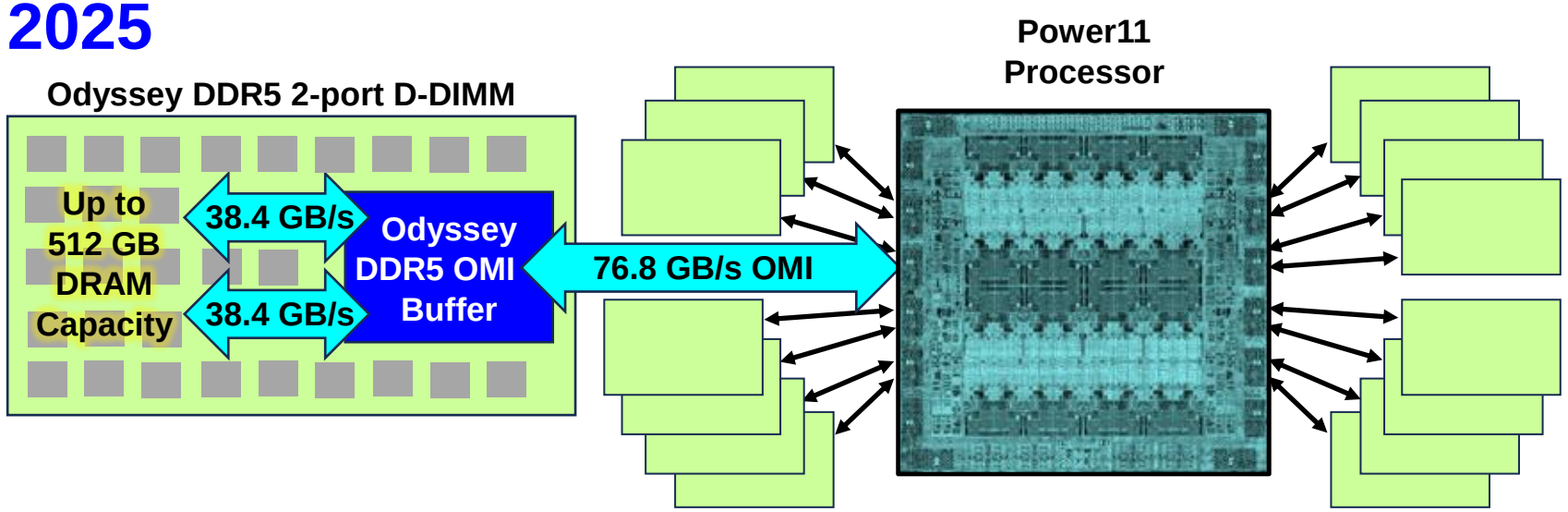


400 GB/s DRAM Bandwidth / Socket

4 Terabyte DRAM Capacity / Socket

700-800 GB/s System Coherence Flow

2025



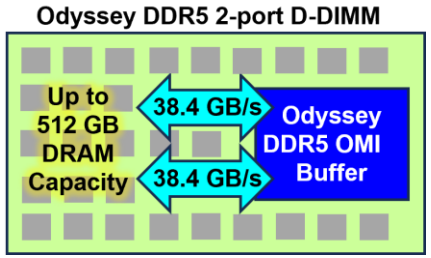
1200 GB/s DRAM Bandwidth / Socket **3x!!!**

8 Terabyte DRAM Capacity / Socket **2x!!**

1000 GB/s System Coherence Flow **1.3x!**

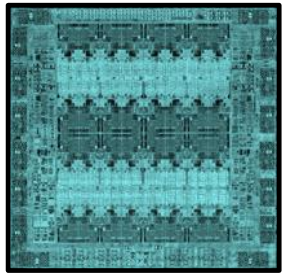
Power11: Full Stack Innovation and Cross-Optimization

Memory Architecture
Energy / Thermal Infrastructure



Agnostic, 3x Pipes, 2x Capacity
Advanced Cooling Technology

Processor Architecture
Socket-level Packaging
Semiconductor Technology



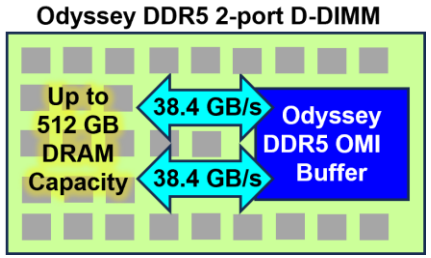
Improved Thread, Core, Capacity
ISC Silicon Layer: Energy Optimization
Samsung Foundries Enhanced 7nm

Power11: Full Stack Innovation and Cross-Optimization

Platform Capabilities

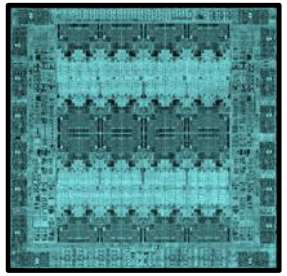
Uptime, Energy Mgmt, Security, Resource Pools

Memory Architecture
Energy / Thermal Infrastructure



Agnostic, 3x Pipes, 2x Capacity
Advanced Cooling Technology

Processor Architecture
Socket-level Packaging
Semiconductor Technology



Improved Thread, Core, Capacity
ISC Silicon Layer: Energy Optimization
Samsung Foundries Enhanced 7nm

Power11: Full Stack Innovation and Cross-Optimization

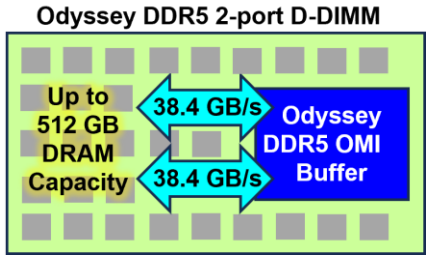
AI Acceleration

Accelerated AI (all systems)
IBM Spyre Accelerator
Optimized for Inference

Platform Capabilities

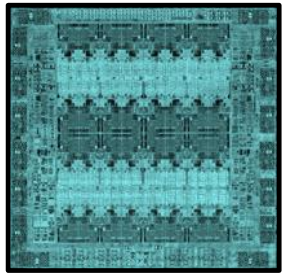
Uptime, Energy Mgmt, Security, Resource Pools

Memory Architecture
Energy / Thermal Infrastructure



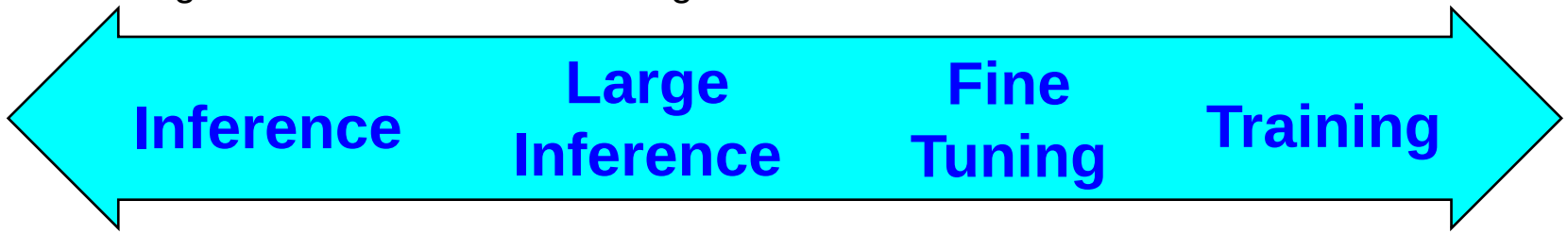
Agnostic, 3x Pipes, 2x Capacity
Advanced Cooling Technology

Processor Architecture
Socket-level Packaging
Semiconductor Technology

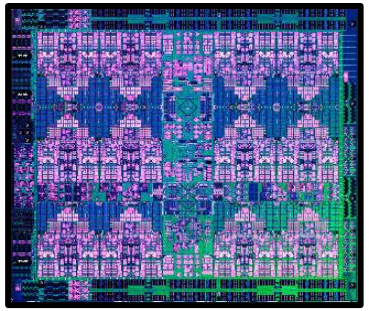


Improved Thread, Core, Capacity
ISC Silicon Layer: Energy Optimization
Samsung Foundries Enhanced 7nm

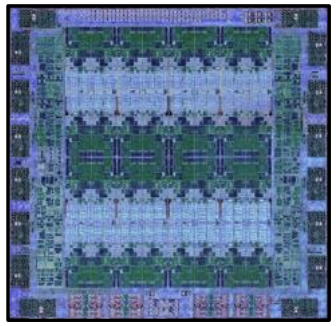
AI Improvements: Accelerated support for Large Models and Model Tuning



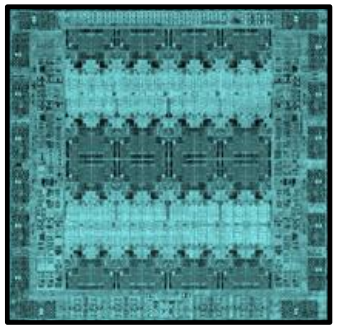
POWER9



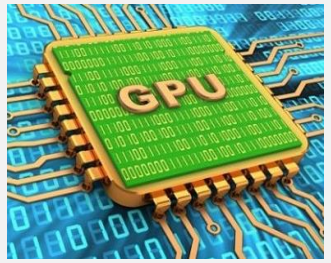
Power10



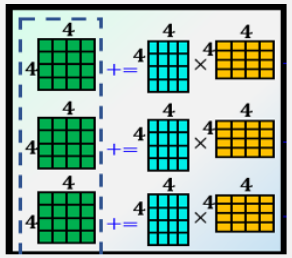
Power11



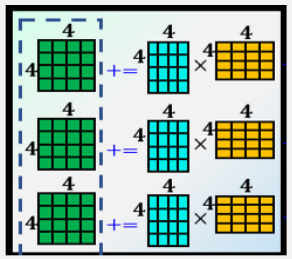
External



In-Core MMA

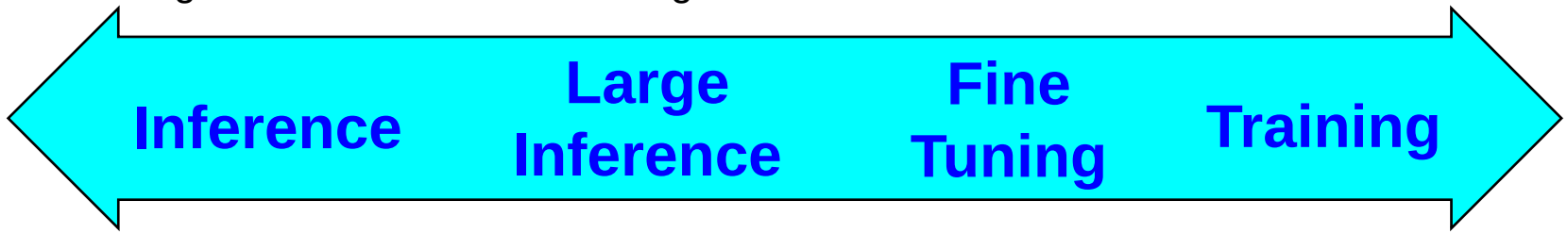


In-Core MMA

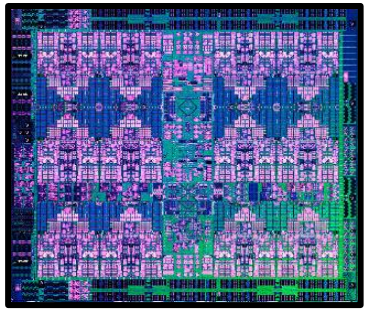


External Spyre Accelerator

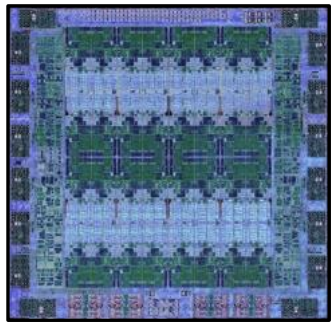
AI Improvements: Accelerated support for Large Models and Model Tuning



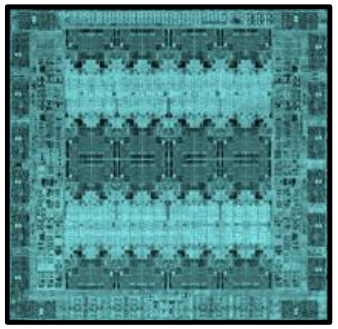
POWER9



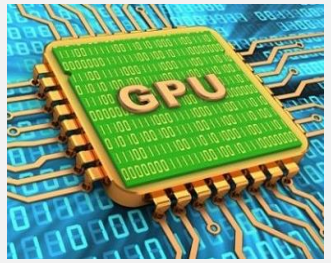
Power10



Power11

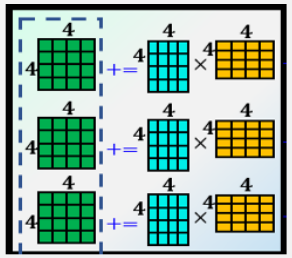


External



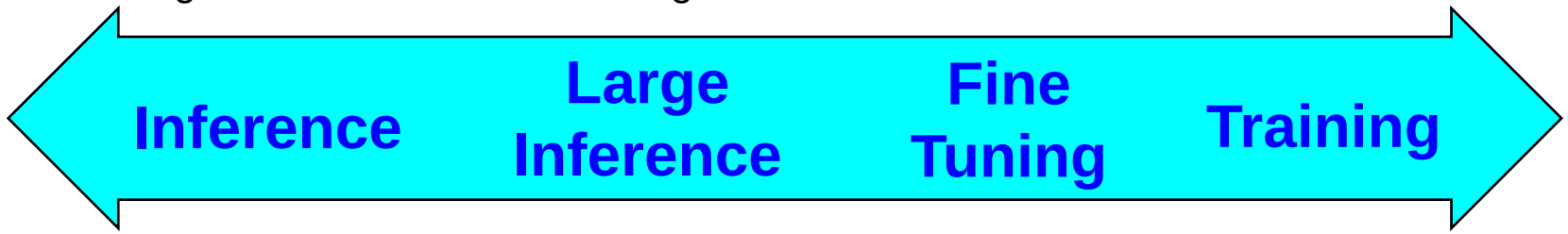
In-Core MMA

In-Core MMA

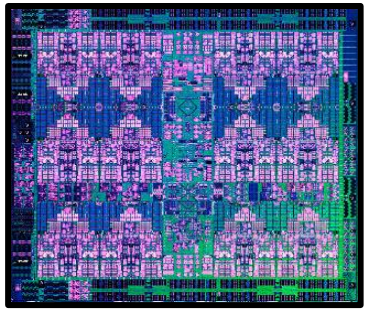


External Spyre Accelerator

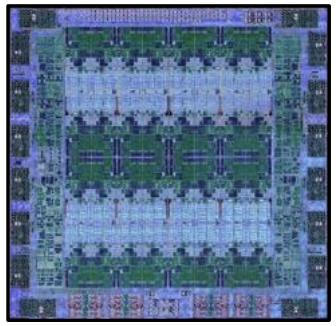
AI Improvements: Accelerated support for Large Models and Model Tuning



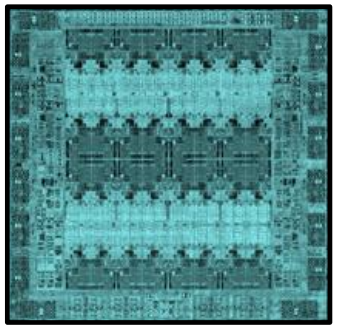
POWER9

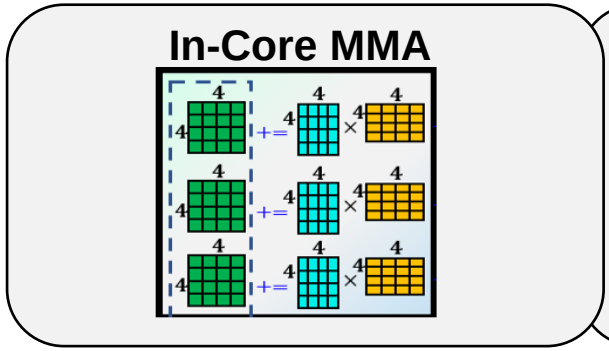


Power10

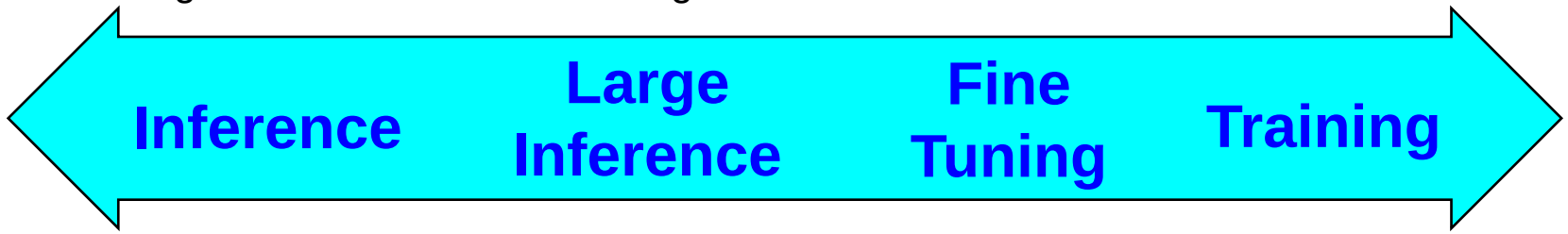


Power11

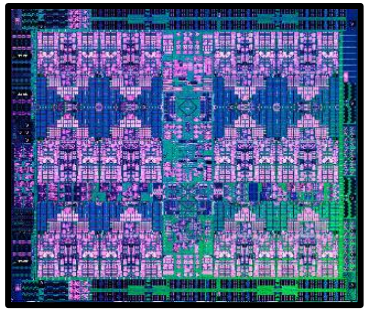




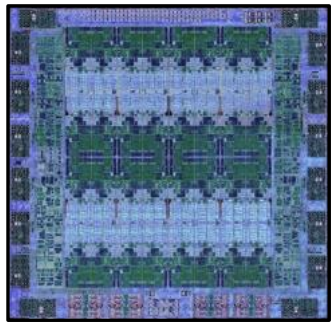
AI Improvements: Accelerated support for Large Models and Model Tuning



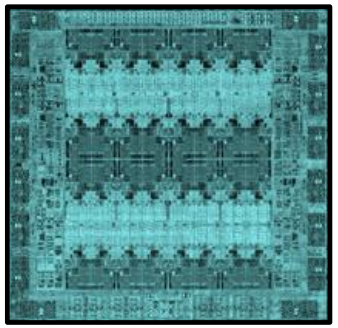
POWER9

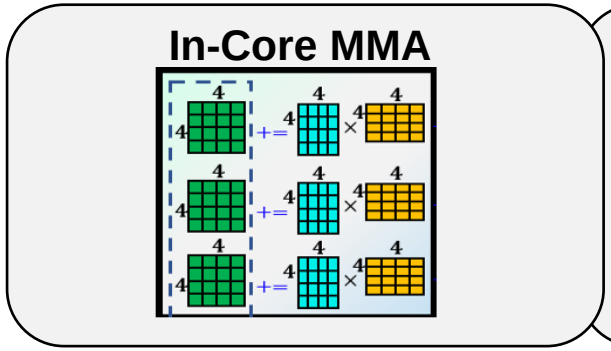


Power10



Power11





Power11: Full Stack Innovation and Cross-Optimization

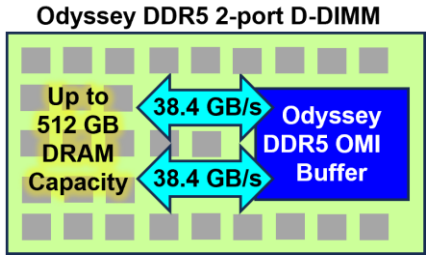
AI Acceleration

Accelerated AI (all systems)
IBM Spyre Accelerator
Optimized for Inference

Platform Capabilities

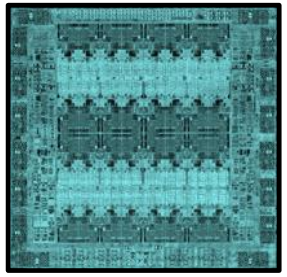
Uptime, Energy Mgmt, Security, Resource Pools

Memory Architecture
Energy / Thermal Infrastructure



Agnostic, 3x Pipes, 2x Capacity
Advanced Cooling Technology

Processor Architecture
Socket-level Packaging
Semiconductor Technology

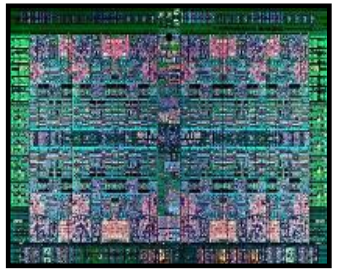


Improved Thread, Core, Capacity
ISC Silicon Layer: Energy Optimization
Samsung Foundries Enhanced 7nm

IBM Power Processor Roadmap

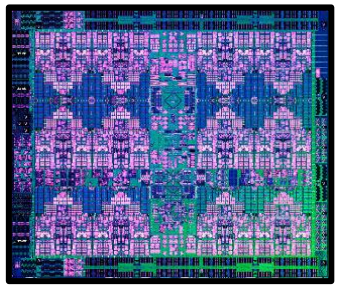
After Power11: Chiplet-based Architecture

POWER8



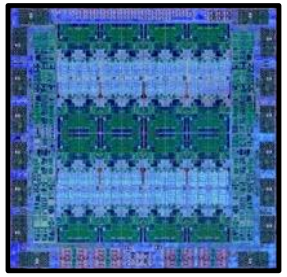
Powerful SMT8 Core
Enterprise Scaling
Big Data Optimized
Agnostic Memory

POWER9



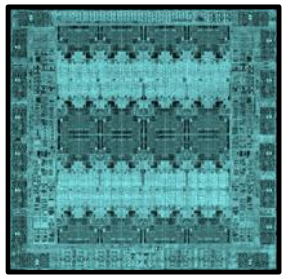
Modular Core Design
Accelerator Attach
(NVlink, OpenCAPI)
Data Plane Bandwidth
DDR & CDIMM Memory

Power10



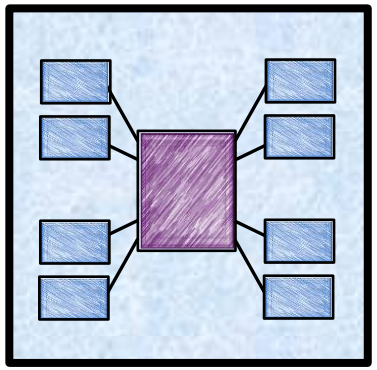
Core / Thread Strength
Socket Performance
In-Core AI Acceleration
Efficiency / Sustainability
HW Accelerated Security

Power11

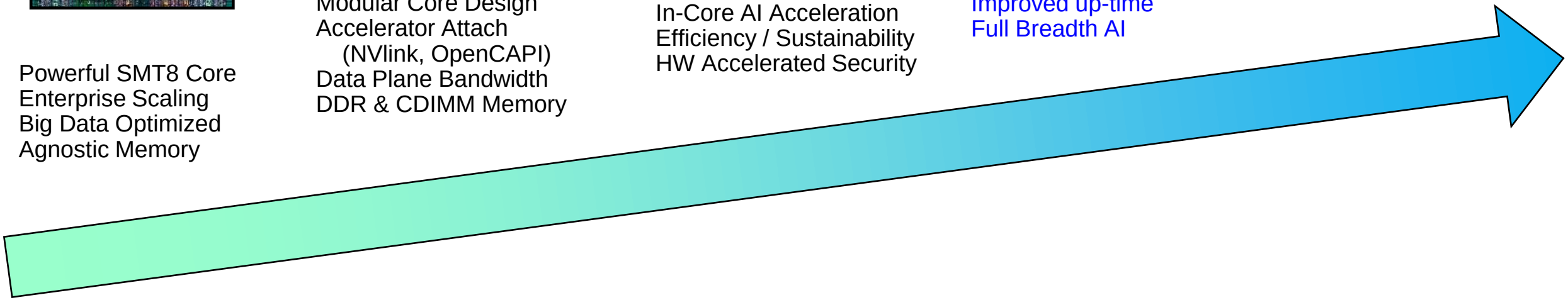


Core / Thread Strength
Socket Performance
Enterprise Scaling
Energy Optimization
Improved up-time
Full Breadth AI

Power Future



(Under development)

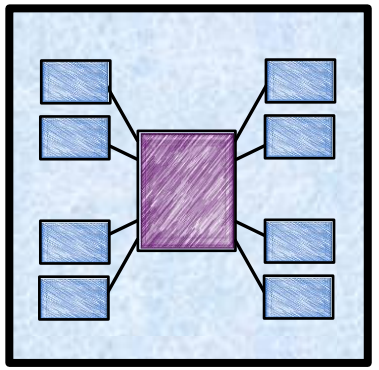


Information is forward looking and plans are subject to change based upon IBM's discretion

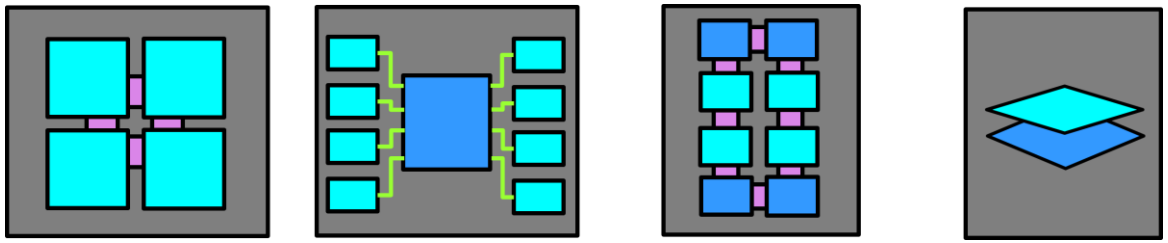
IBM Power Processor Roadmap

After Power11: Chiplet-based Architecture

Power Future



- Moore’s Law Challenge: Silicon Scaling is Slowing
- Industry shift: Leverage Packaging: Chiplets, 3D

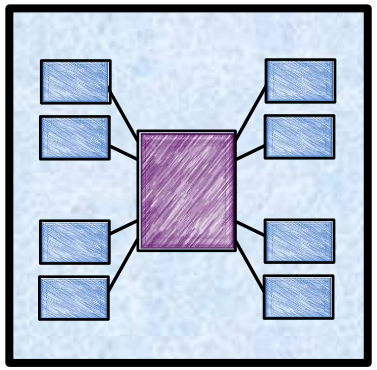


- Several years back we assessed for Power …

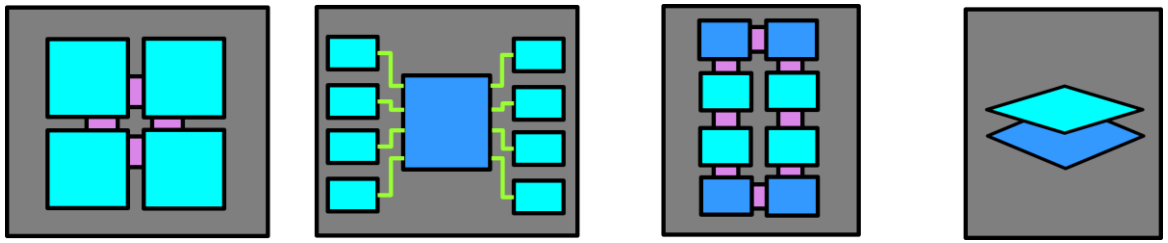
IBM Power Processor Roadmap

After Power11: Chiplet-based Architecture

Power Future



- **Moore's Law Challenge: Silicon Scaling is Slowing**
- **Industry shift: Leverage Packaging: Chiplets, 3D**



Strong Value for Power Roadmap: Multiple generations into Future

1) 3x Silicon per Socket (small chiplets with room to grow)
2) Manufacturing Yield Synergies
3) Ability to maintain Strong Bandwidth across Chiplets
4) OMI Memory Beachfront Efficiency → Enable High Bandwidth SMP/IO Interfaces
5) Substantial Latency Reduction / Topology Synergy: Continued Robust Scaling
6) Long-term Development Efficiency and Flexibility