IBM’s Power11 Processor

Hot Chips 2025
August 24-26, 2025

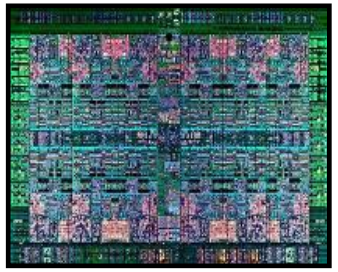
William Starke

IBM Power is

- Mission Critical Reliability & Security
- Robust Scaling: Small Business to World’s Largest Enterprises
- Massive Data Centric / Highly Virtualized
- Flexible Deployment: On-prem, Hybrid, Public Cloud
- Workload Modernization / AI Infusion
- System-Focused, not Socket-Focused

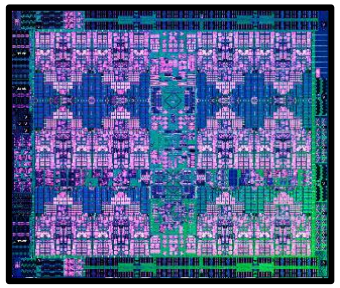
IBM Power Processor Roadmap

POWER8



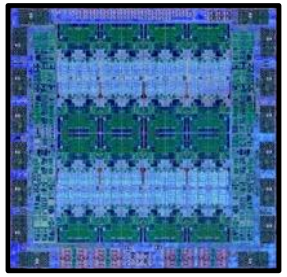
Powerful SMT8 Core
Enterprise Scaling
Big Data Optimized
Agnostic Memory

POWER9



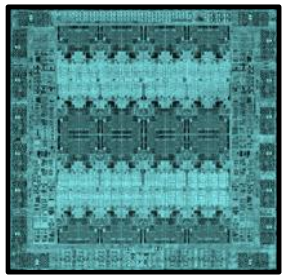
Modular Core Design
Accelerator Attach
(NVlink, OpenCAPI)
Data Plane Bandwidth
DDR & CDIMM Memory

Power10



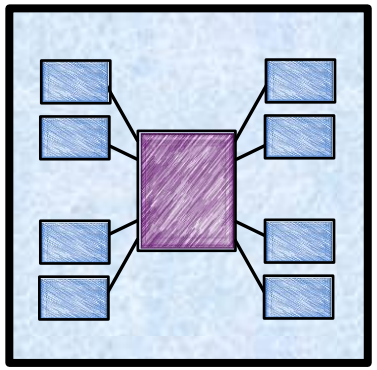
Core / Thread Strength
Socket Performance
In-Core AI Acceleration
Efficiency / Sustainability
HW Accelerated Security

Power11



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

Power Future



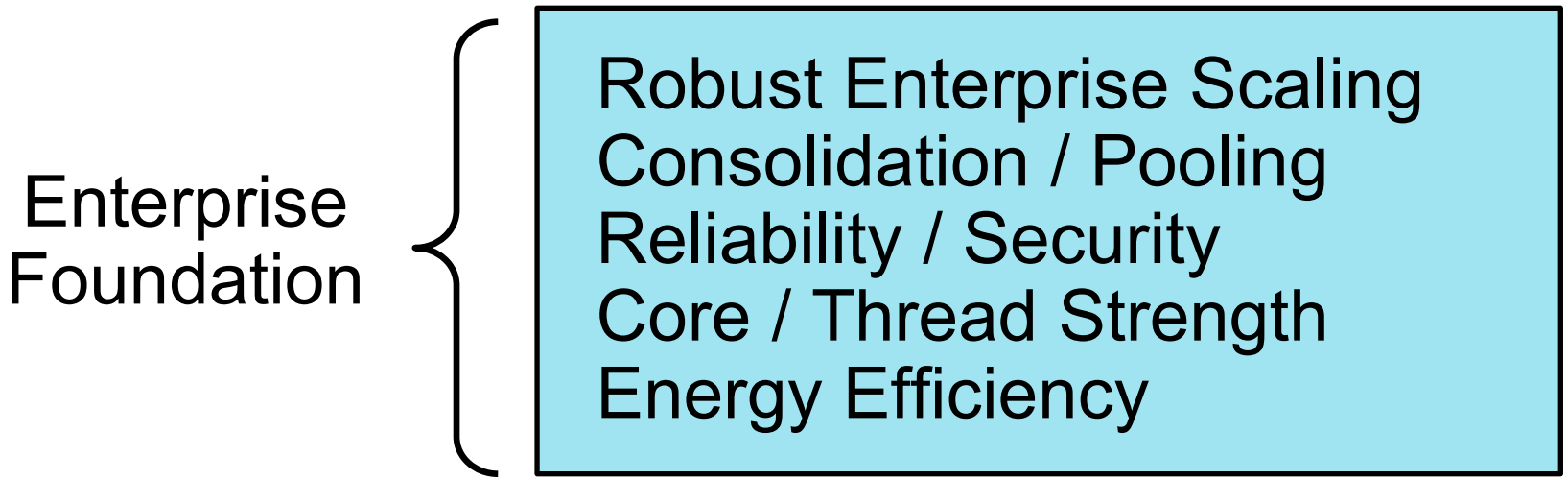
(Under development)

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

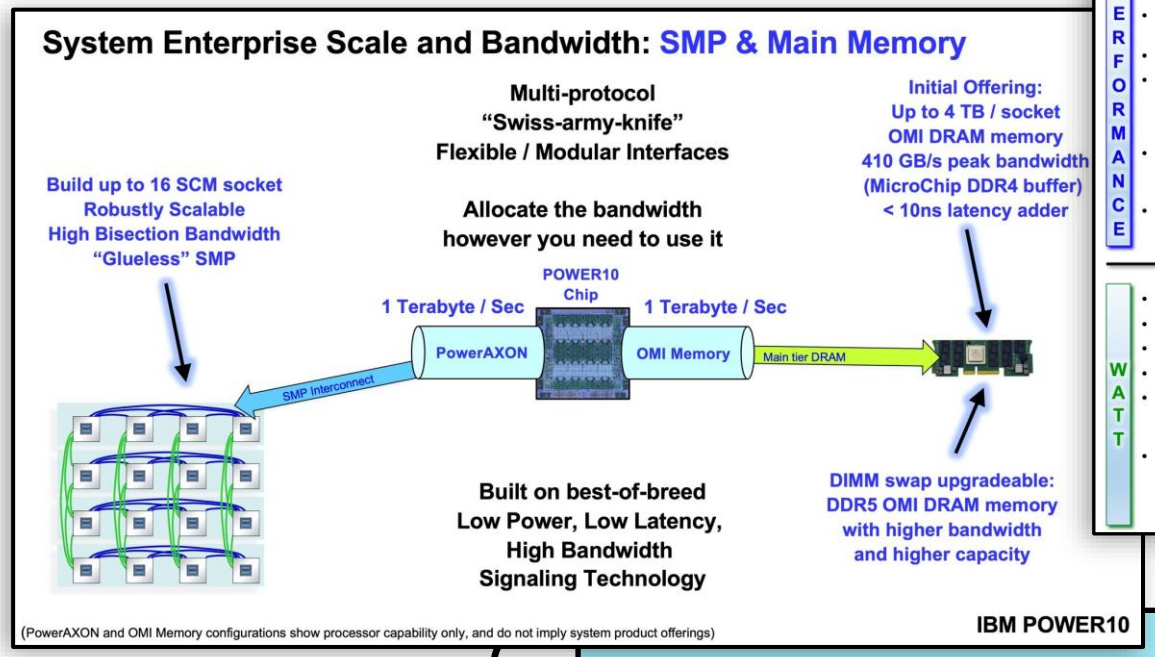
IBM Power11 Product Launch

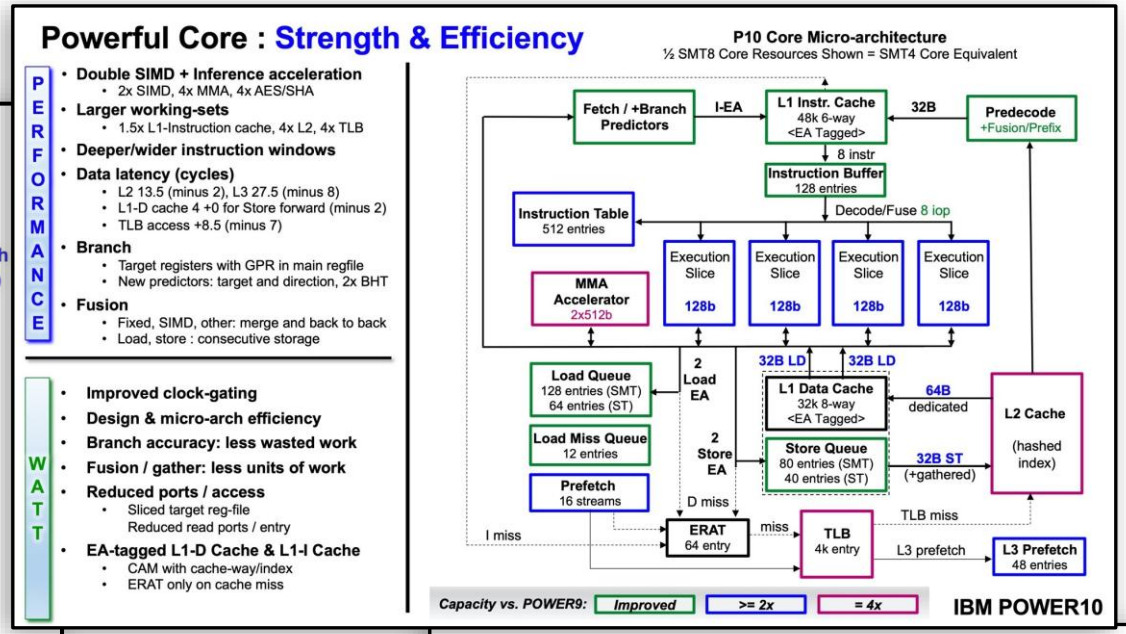
From Power10 to Power11…

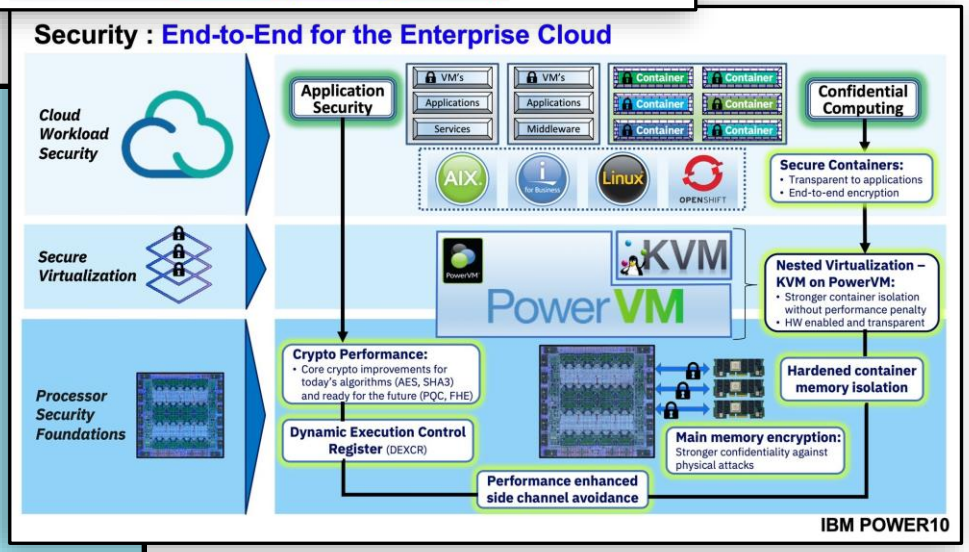
Power10 Architecture:
- Tremendously Successful



From Power10 to Power11…







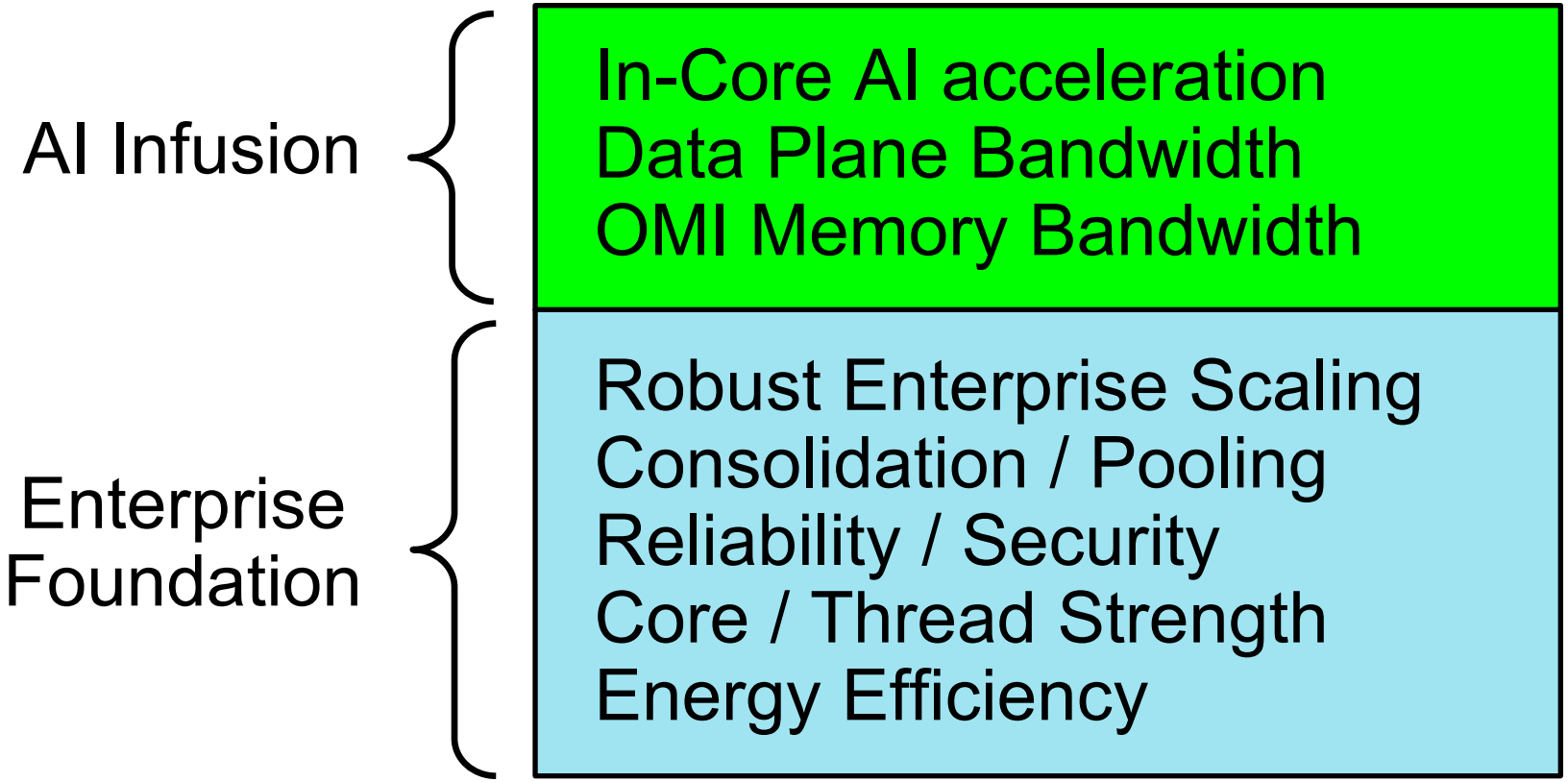
Enterprise Foundation

Robust Enterprise Scaling
Consolidation / Pooling
Reliability / Security
Core / Thread Strength
Energy Efficiency

From Power10 to Power11…

Power10 Architecture:

- Tremendously Successful
- Ahead of its time for AI Infusion



From Power10 to Power11…

Power10 Architecture:

- Tremendously Successful
- Ahead of its time for AI Infusion

AI Infusion

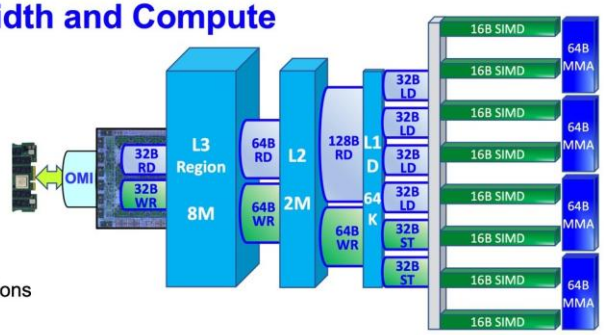
In-Core AI acceleration
Data Plane Bandwidth
OMI Memory Bandwidth

Powerful Core : AI Infused Bandwidth and Compute

BYTES

2x Bytes from all sources
(OMI, L3, L2, L1 caches*)

- **4 32B loads, 2 32B stores per SMT8 Core**
  - New ISA or fusion
  - Thread max 2 32B loads, 1 32B store
- **OMI Memory to one Core**
  - 256 GB/s peak, 120 GB/s sustained
  - With 3x L3 prefetch and memory prefetch extensions



FLOP

2x Bandwidth matched SIMD*

- **8 independent SIMD engines per Core**
  - Fixed, float, permute

4-32x Matrix Math Acceleration*

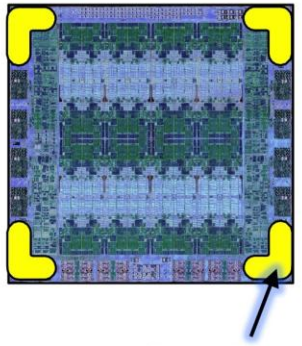
- **4 512b engines per core = 2048b results / cycle**
  - Matrix math outer products: $A \leftarrow \{\pm\}A\ \{\pm\}\ XY^T$
  - Double, Single, Reduced precision

Rank	Operand Type (X,Y)			Accumulator	Peak [FL]OPS / cycle		
k	Type	X	Y^T	A	Instruction	Thread	SMT8 Core
1	Float-64 DP	4×1	1×2	4×2 (Fp-64)	16	32	64
	Float-32 SP	4×1	1×4		32	64	128
2	Float-16 HP	4×2	2×4	4×4 (Fp-32)	64	128	256
	Bfloat-16 HP	4×2	2×4				
	Int-16	4×2	2×4				
4	Int-8	4×4	4×4	4×4 (Int-32)	128	256	512
8	Int 4	4×8	8×4		256	512	1024

* versus POWER9

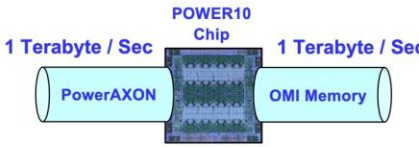
IBM POWER10

System Composability: PowerAXON & Open Memory Interfaces

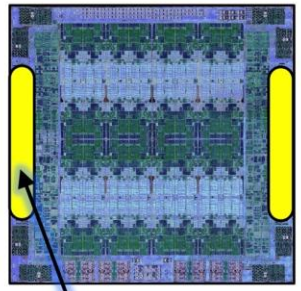


PowerAXON corner
4x8 @ 32 GT/s

Multi-protocol
"Swiss-army-knife"
Flexible / Modular Interfaces



Built on best-of-breed
Low Power, Low Latency,
High Bandwidth
Signaling Technology



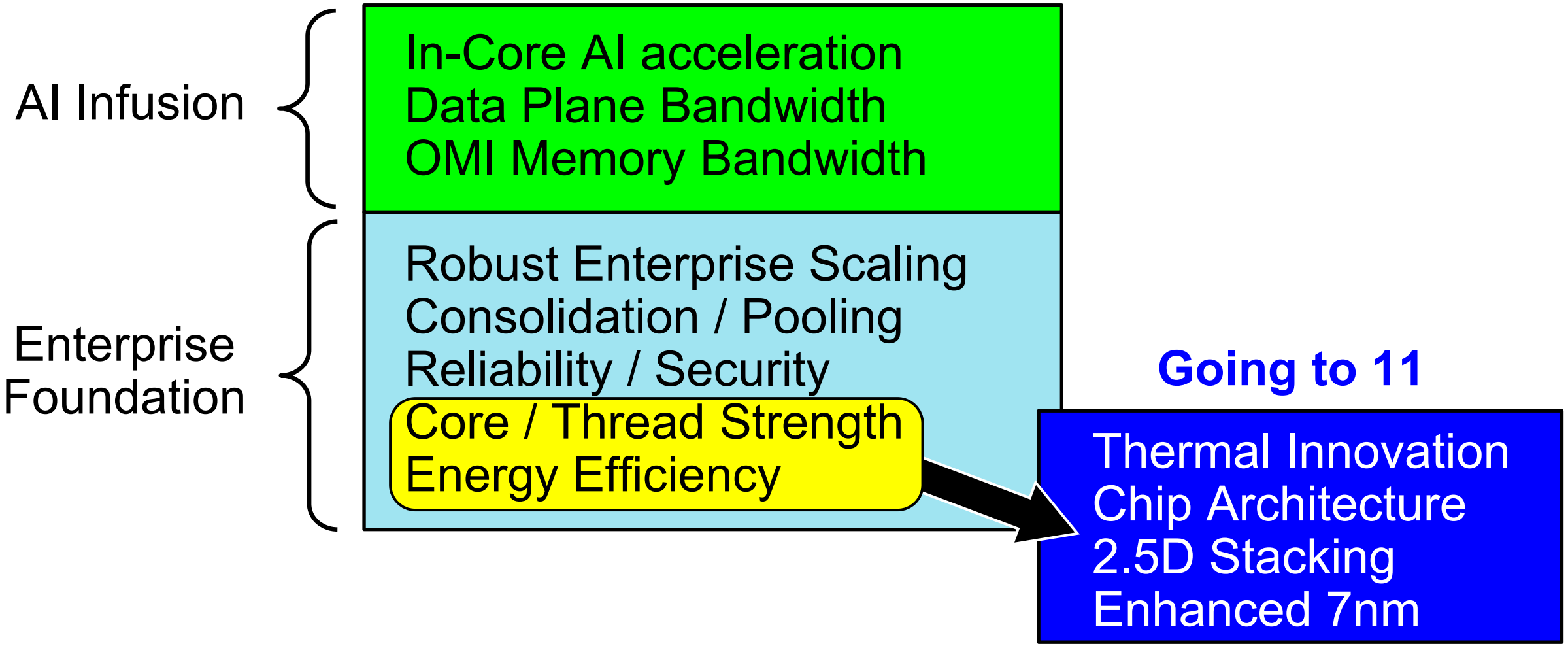
OMI edge
8x8 @ 32 GT/s

6x bandwidth / mm² compared to DDR4 signaling

IBM POWER10

From Power10 to Power11…

Power10 Architecture:
- Tremendously Successful
- Ahead of its time for AI Infusion

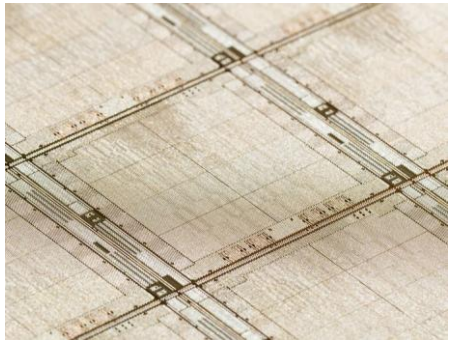


From Power10 to Power11…

2-Socket System: 50% More Cores @ Higher Clock Speed

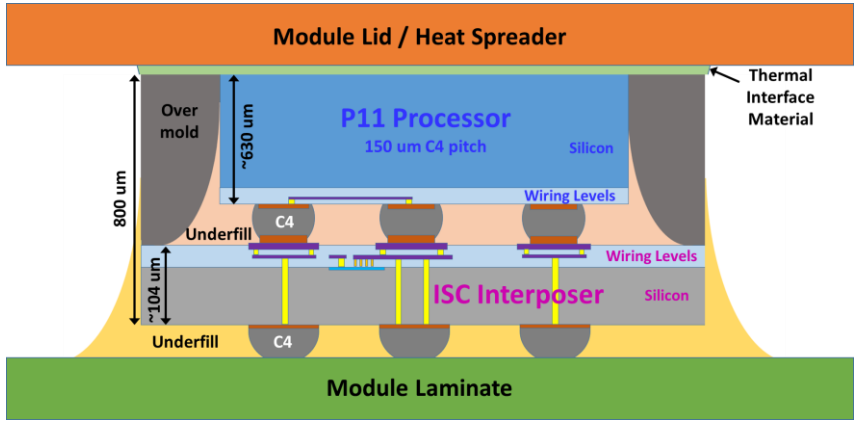
Largest System: 4.0 GHz ➡ 4.3 GHz (with More Cores)

Density vs **Speed**



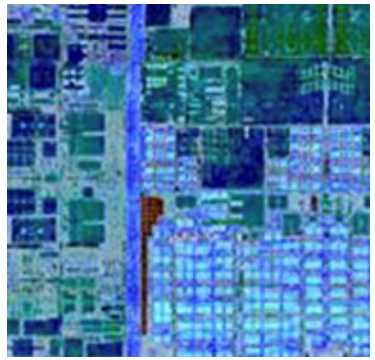
Samsung Foundries
5nm vs **Enhanced 7nm**?

2.5D ISC Architecture

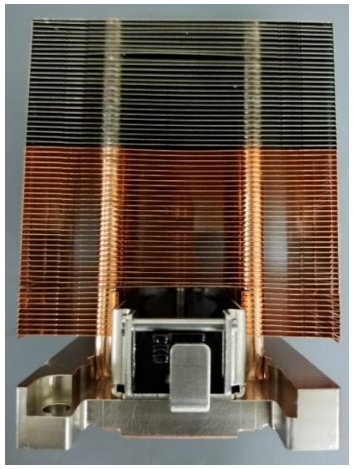


Samsung iCube Si Interposer technology

Chip Architecture



System Thermal Innovation

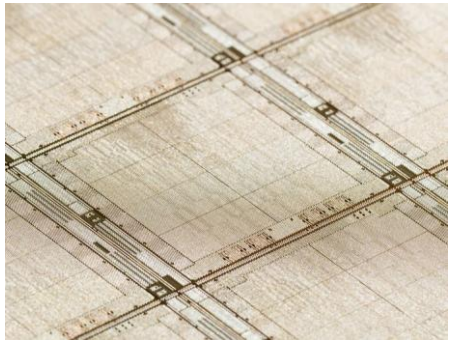


From Power10 to Power11… Relative Performance Growth (rPerf)

(rPerf score is based on weighted composite of multiple enterprise workloads: IBM's formal means of communicating capacity planning data to Power Systems clients)

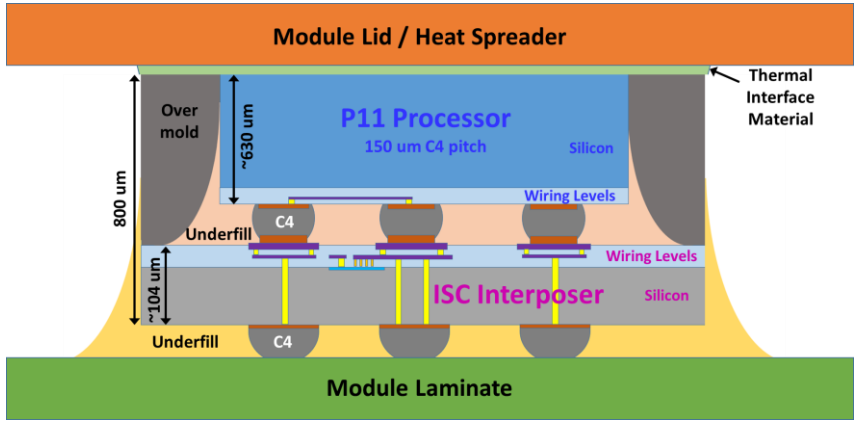
System	Sockets	Form Factor	Power10	Power11	Growth
S1122	2 DCMs	19" x 2U	1024	1532	50%
S1124	2 DCMs	19" x 4U	1331	1737	31%
E1150	4 DCMs	19" x 4U	2689	3504	30%
E1180	16 SCMs	19" x 22U	7999	9141	14%

Density vs **Speed**



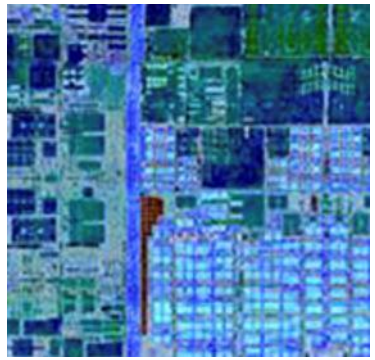
Samsung Foundries 5nm vs **Enhanced 7nm**?

2.5D ISC Architecture

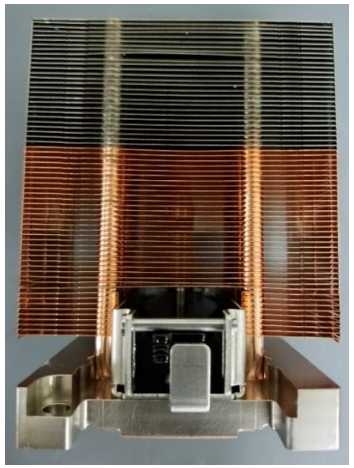


Samsung iCube Si Interposer technology

Chip Architecture

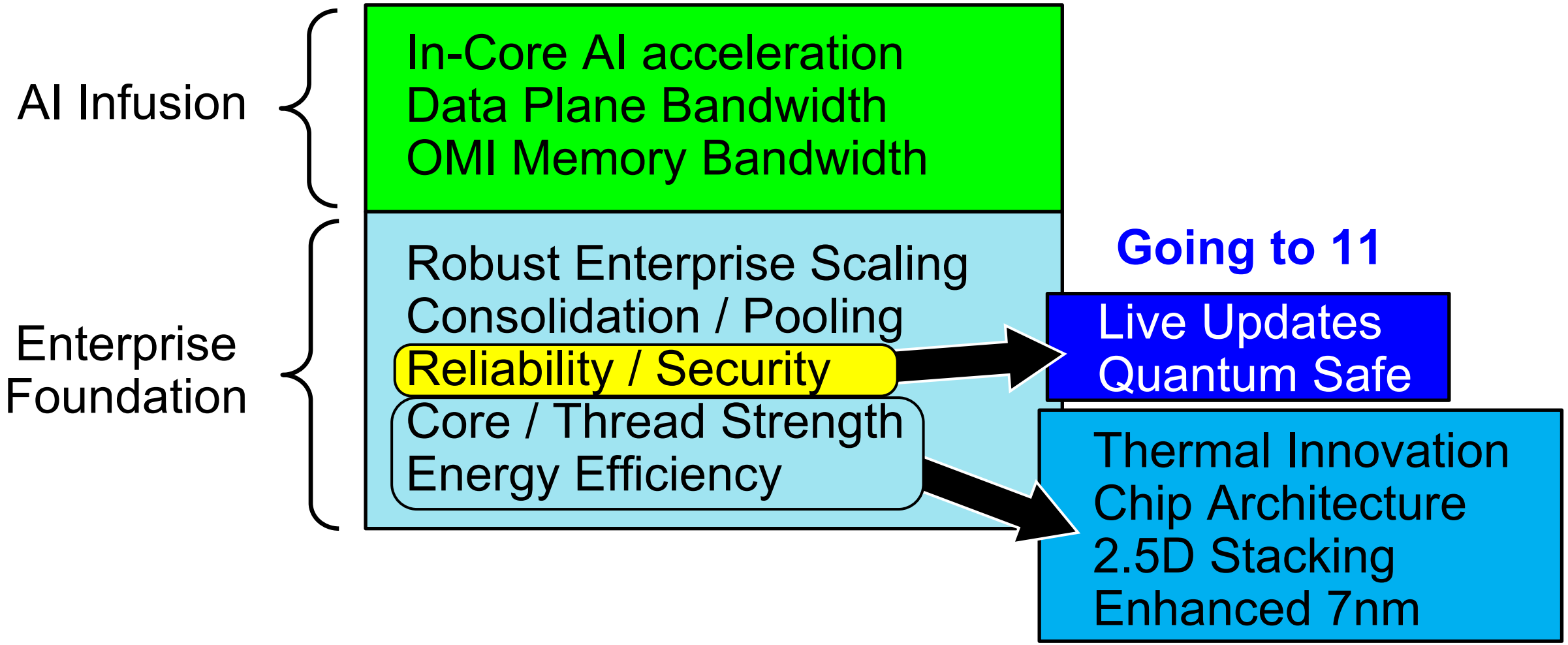


System Thermal Innovation



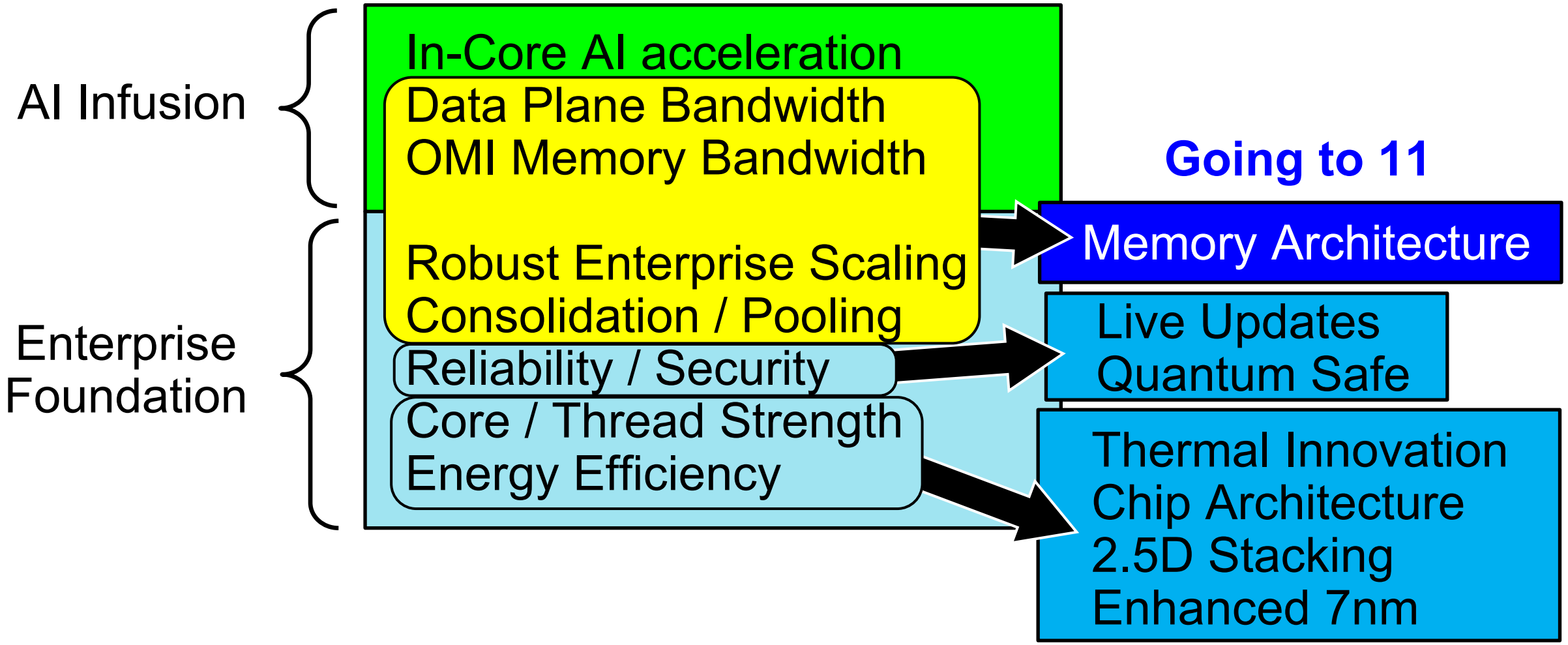
From Power10 to Power11…

Power10 Architecture:
- Tremendously Successful
- Ahead of its time for AI Infusion

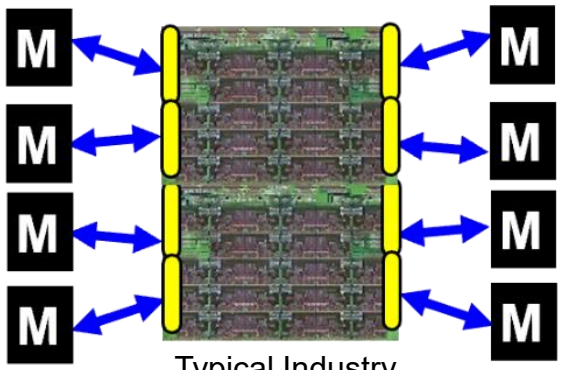


From Power10 to Power11…

Power10 Architecture:
- Tremendously Successful
- Ahead of its time for AI Infusion

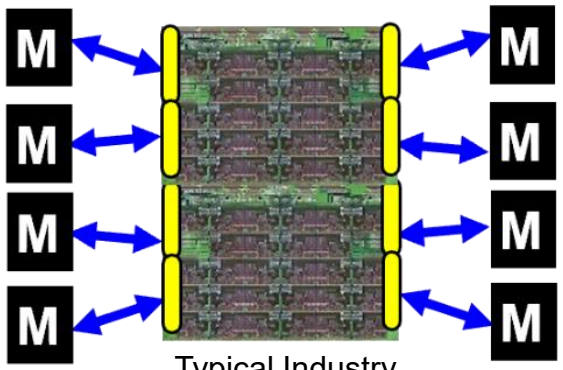


From Power10 to Power11… Strength of OMI Memory Architecture

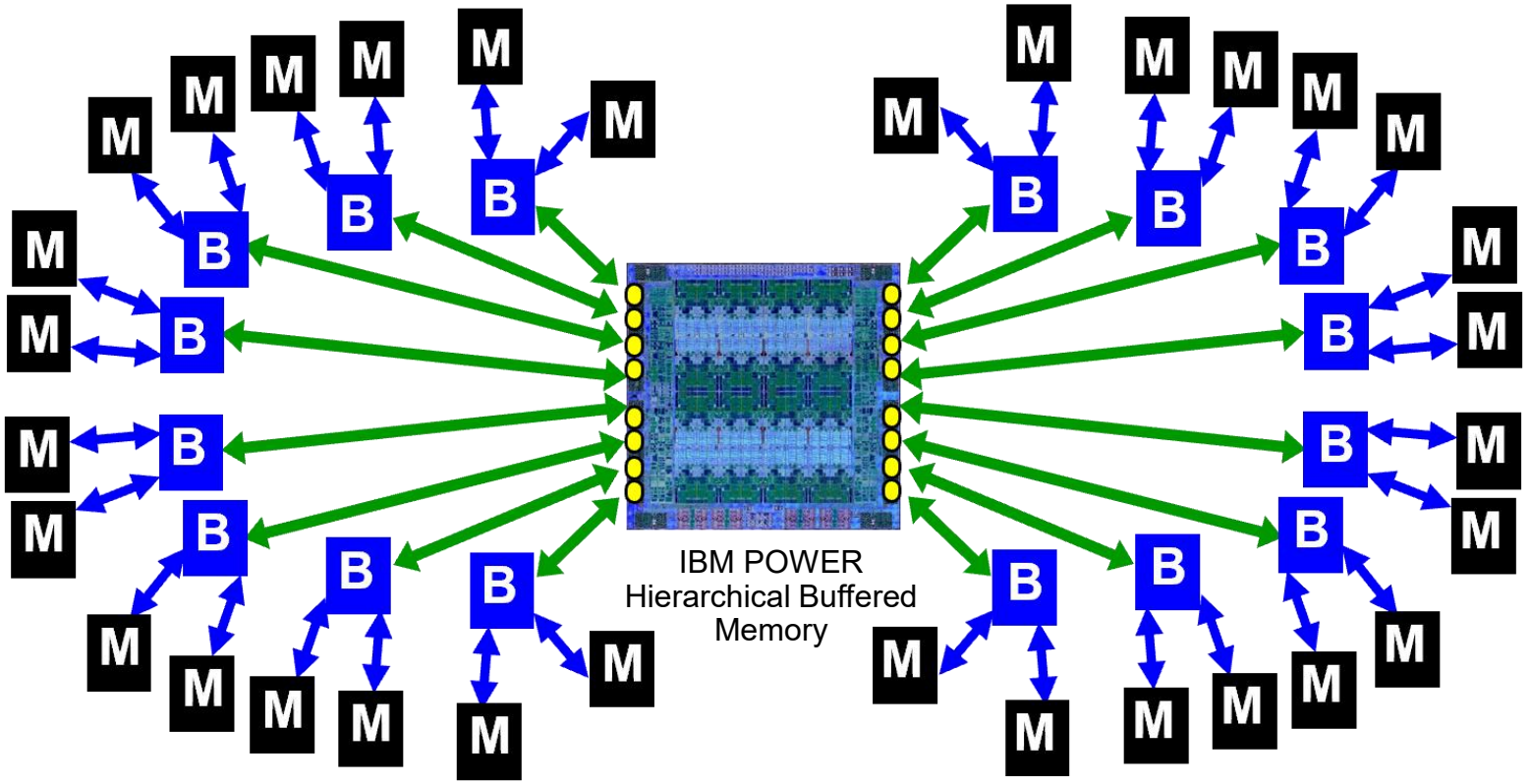


Typical Industry
Direct Attach Memory

From Power10 to Power11… Strength of OMI Memory Architecture

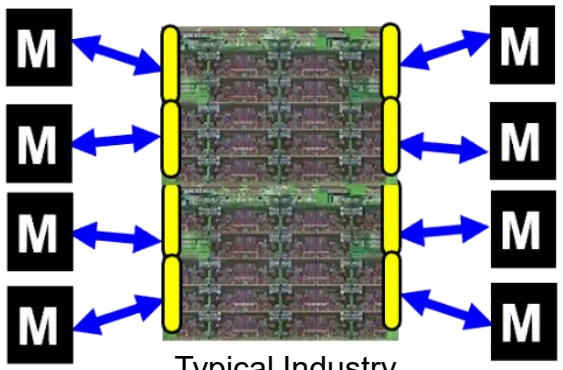


Typical Industry
Direct Attach Memory



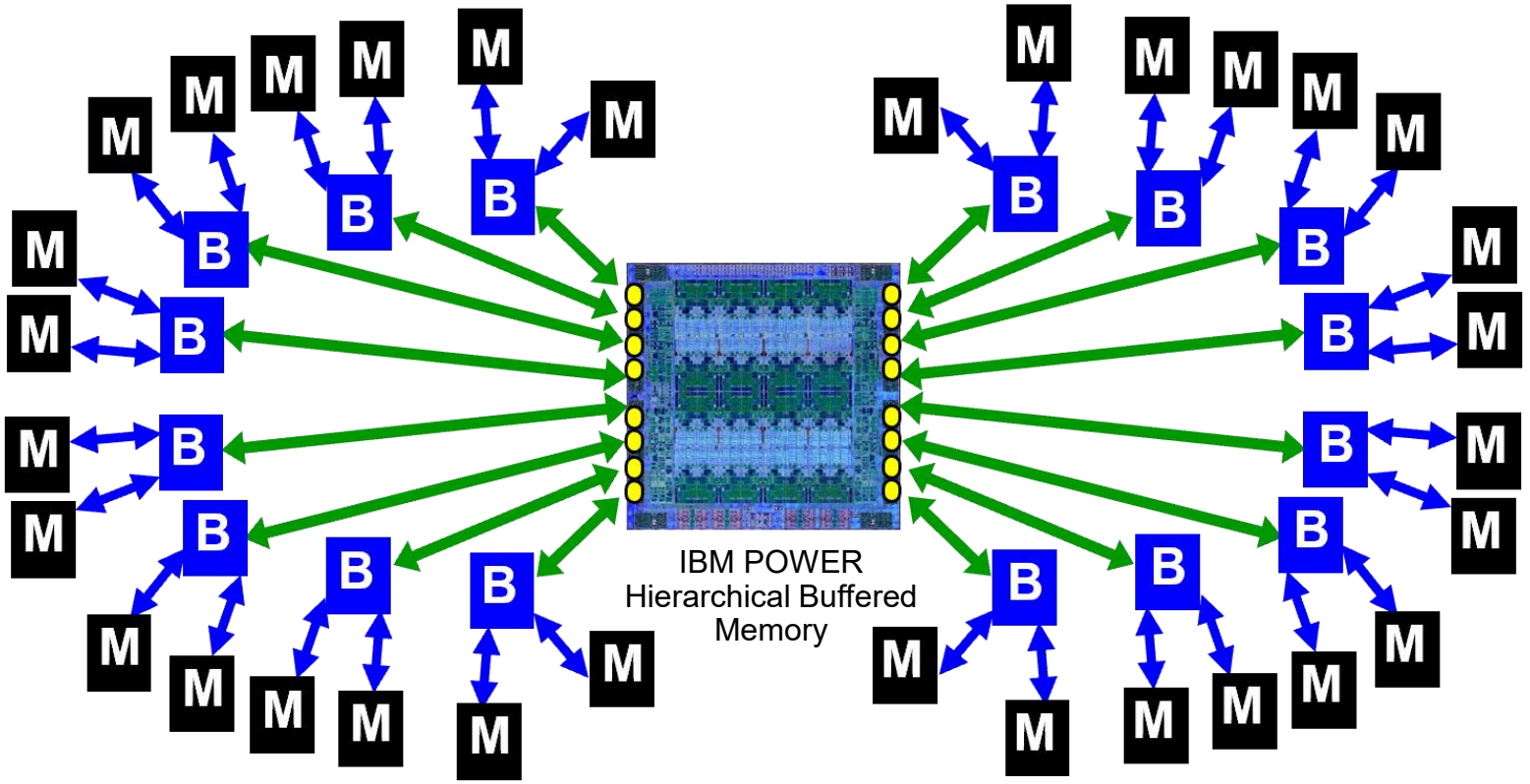
IBM POWER
Hierarchical Buffered
Memory

From Power10 to Power11… Strength of OMI Memory Architecture



Typical Industry
Direct Attach Memory

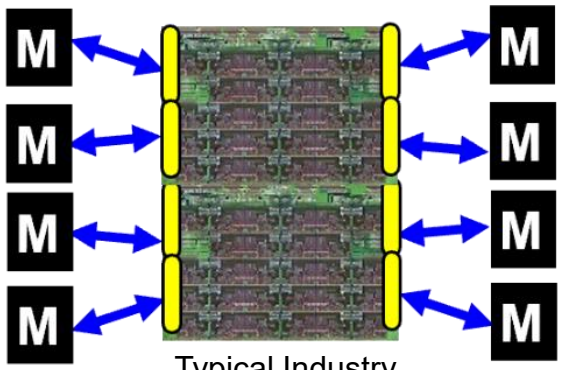
4x Bandwidth / Processor!
4x Capacity / Processor!



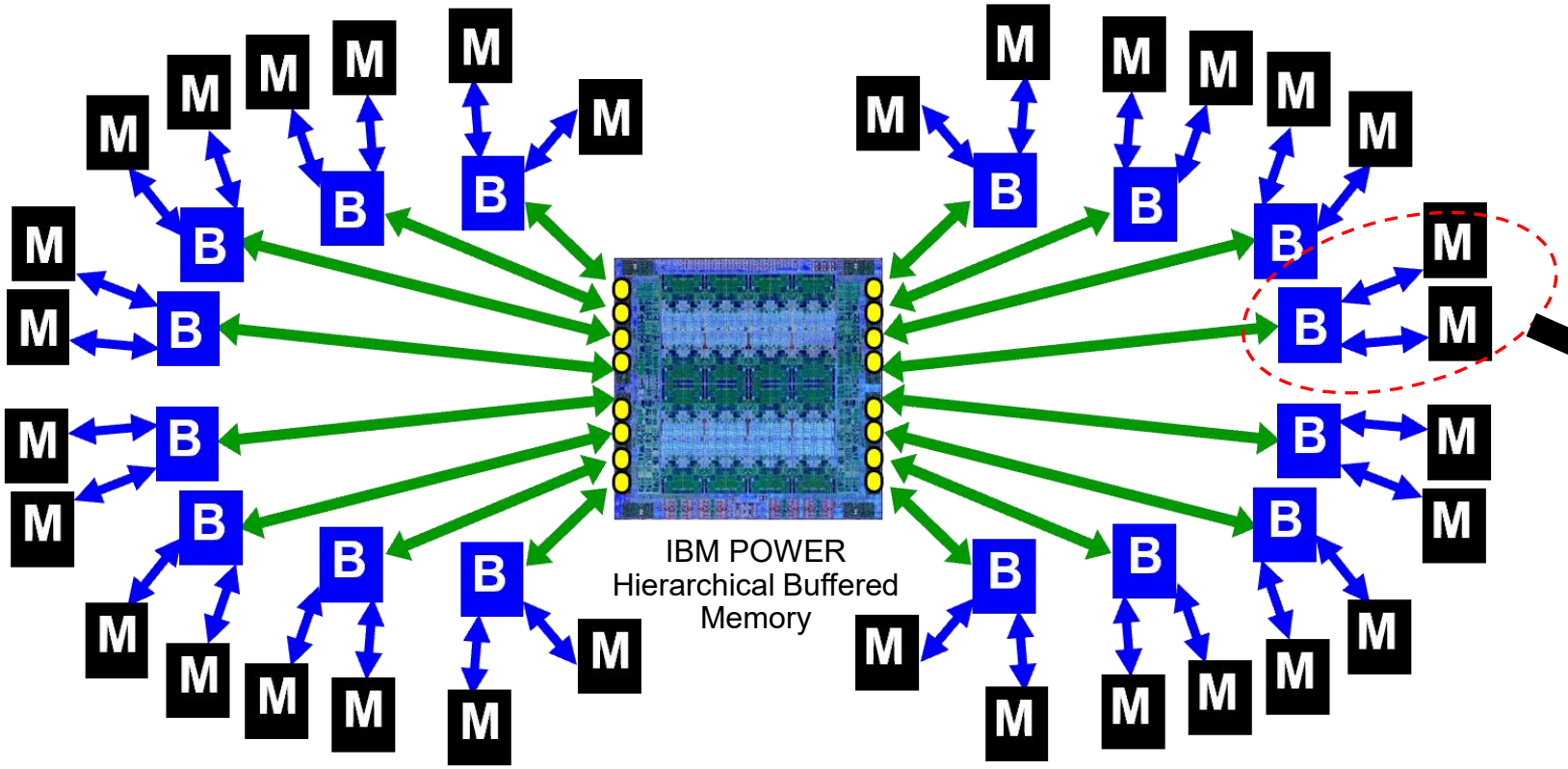
IBM POWER
Hierarchical Buffered
Memory

From Power10 to Power11… Strength of OMI Memory Architecture

4x Bandwidth / Processor!
4x Capacity / Processor!



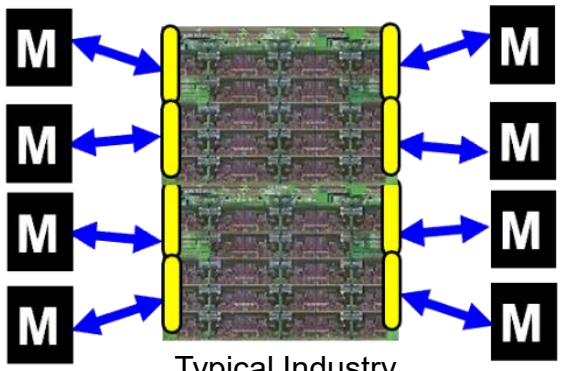
Typical Industry
Direct Attach Memory



IBM POWER
Hierarchical Buffered
Memory

OMI D-DIMM

From Power10 to Power11… Strength of OMI Memory Architecture

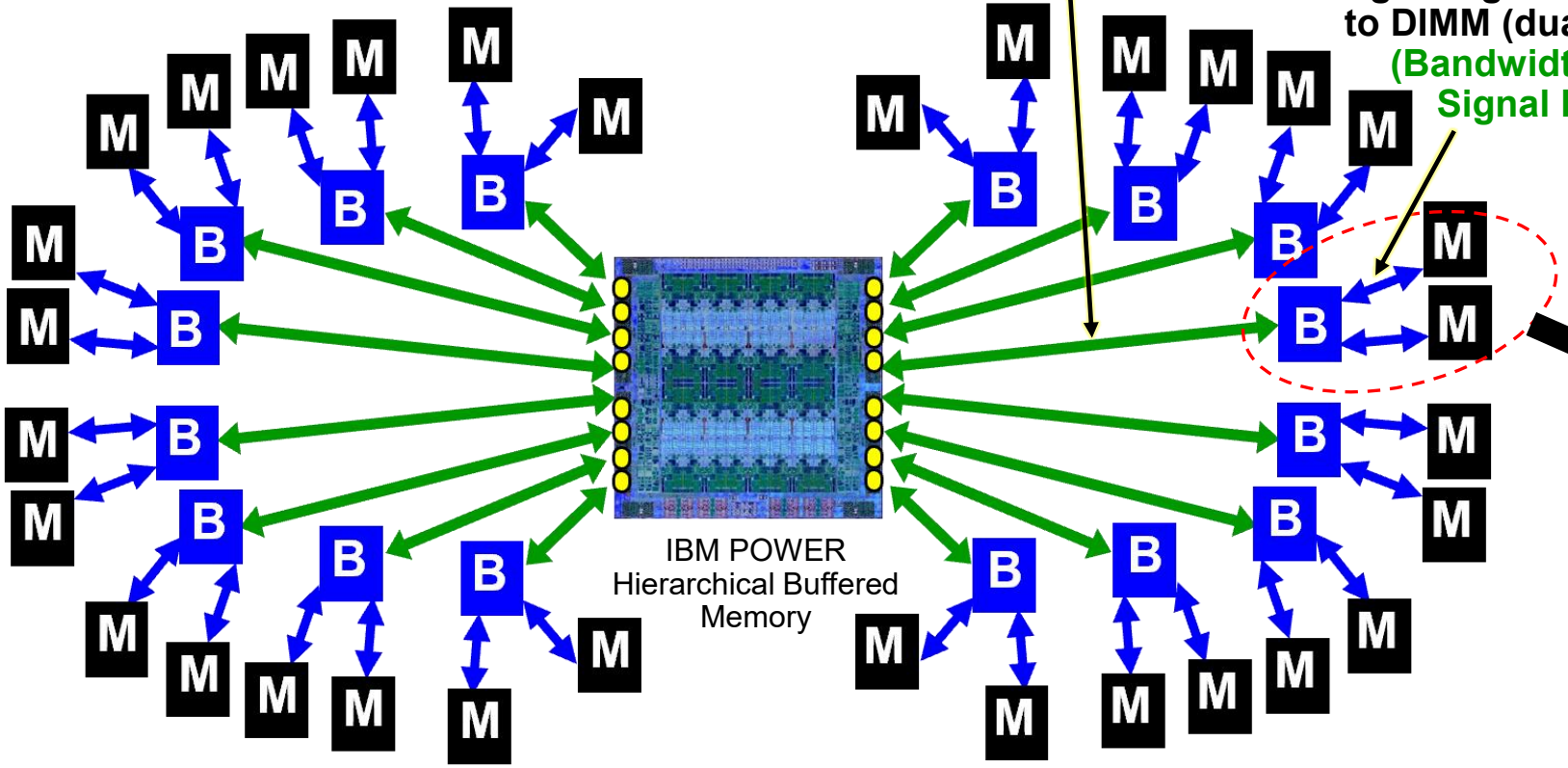


Typical Industry
Direct Attach Memory

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!

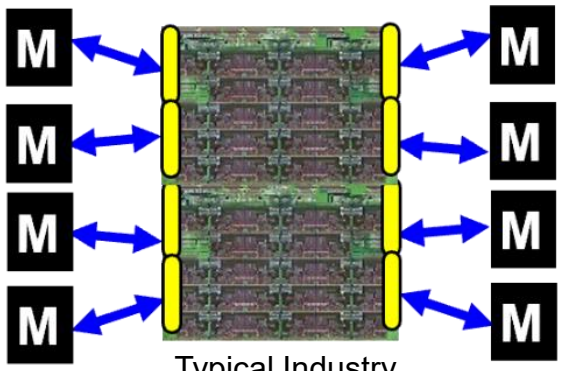
DDR DRAM single-ended signaling running contained to DIMM (dual port fan-out)
(Bandwidth, Capacity, Signal Integrity!)



IBM POWER
Hierarchical Buffered
Memory

OMI D-DIMM

From Power10 to Power11… Strength of OMI Memory Architecture

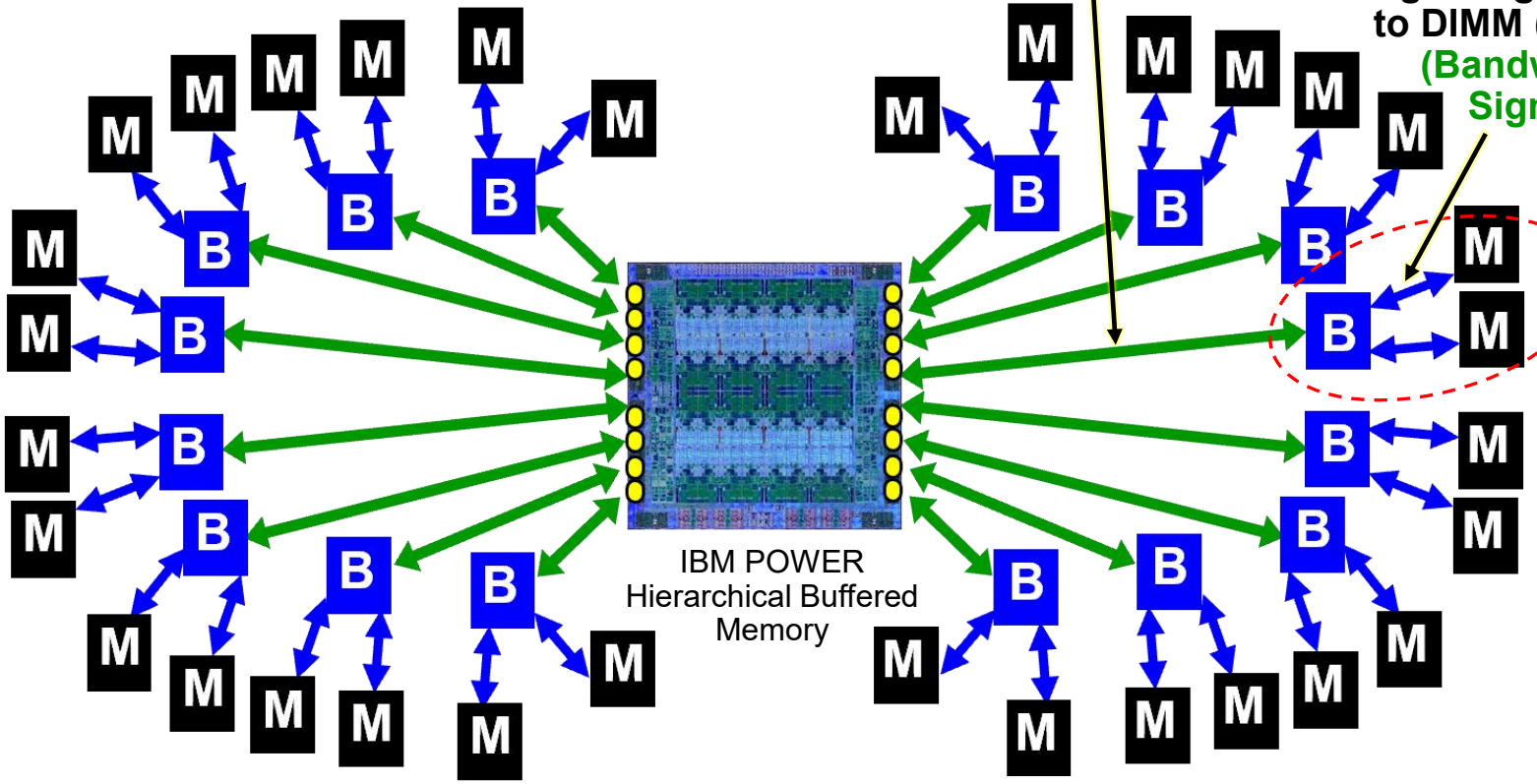


Typical Industry
Direct Attach Memory

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!

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



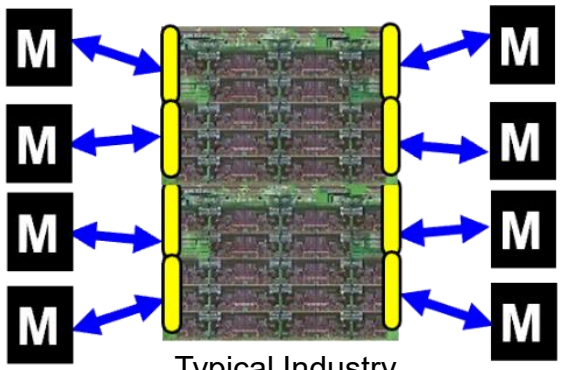
IBM POWER
Hierarchical Buffered
Memory

Spare DRAMs
(Reliability!)

DIMM Connector
(Protocol provides Resiliency!)

OMI D-DIMM

From Power10 to Power11… Strength of OMI Memory Architecture



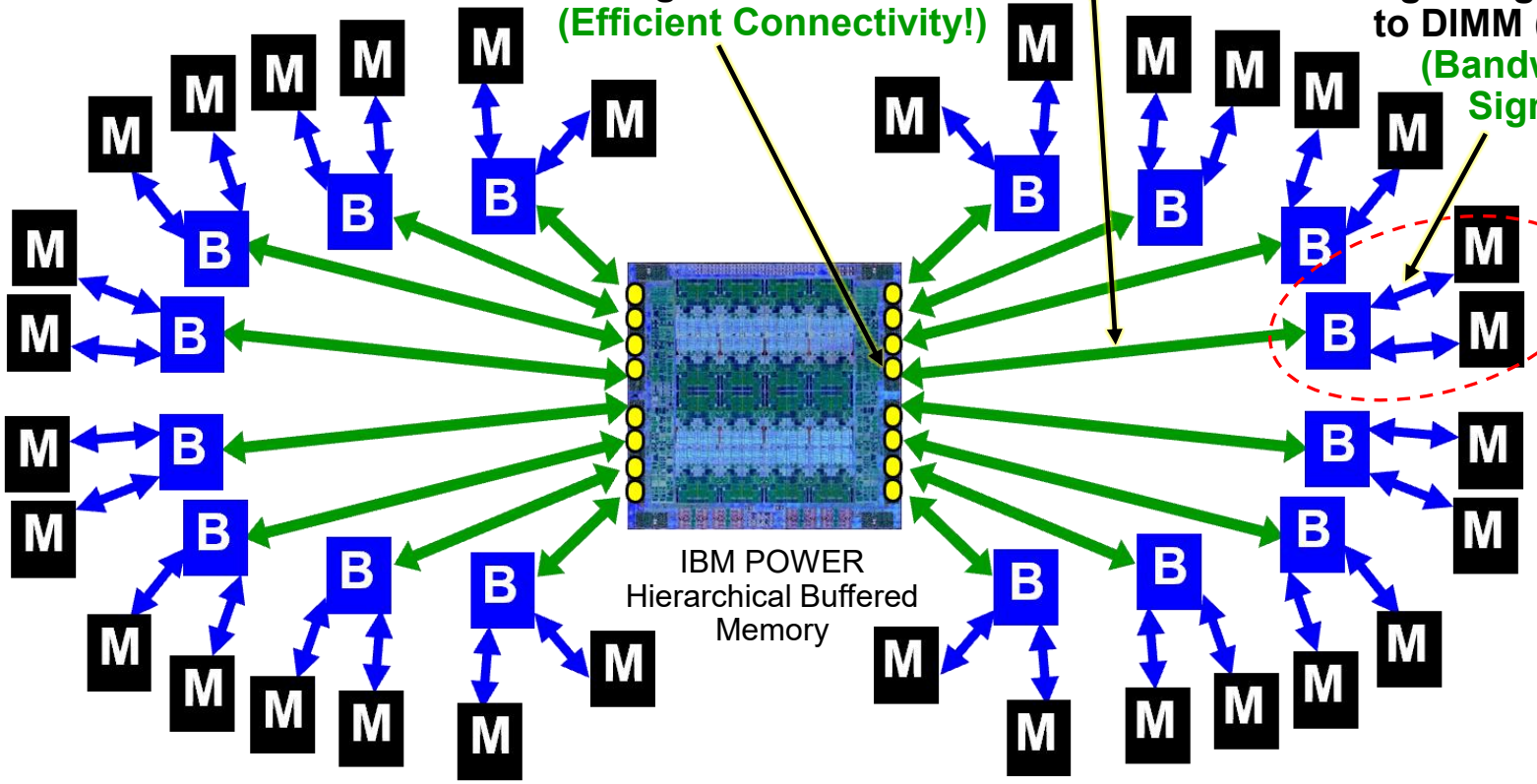
Typical Industry
Direct Attach Memory

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!

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

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



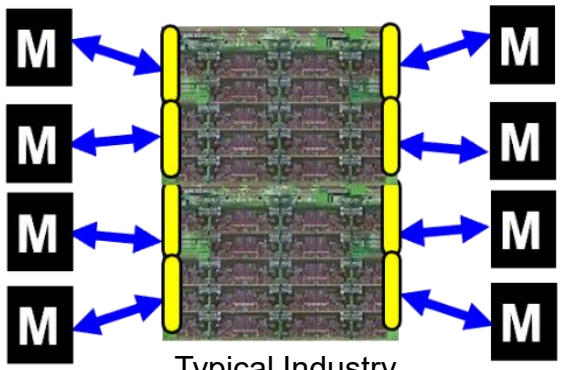
IBM POWER
Hierarchical Buffered
Memory

Spare DRAMs
(Reliability!)

DIMM Connector
(Protocol provides Resiliency!)

OMI D-DIMM

From Power10 to Power11… Strength of OMI Memory Architecture



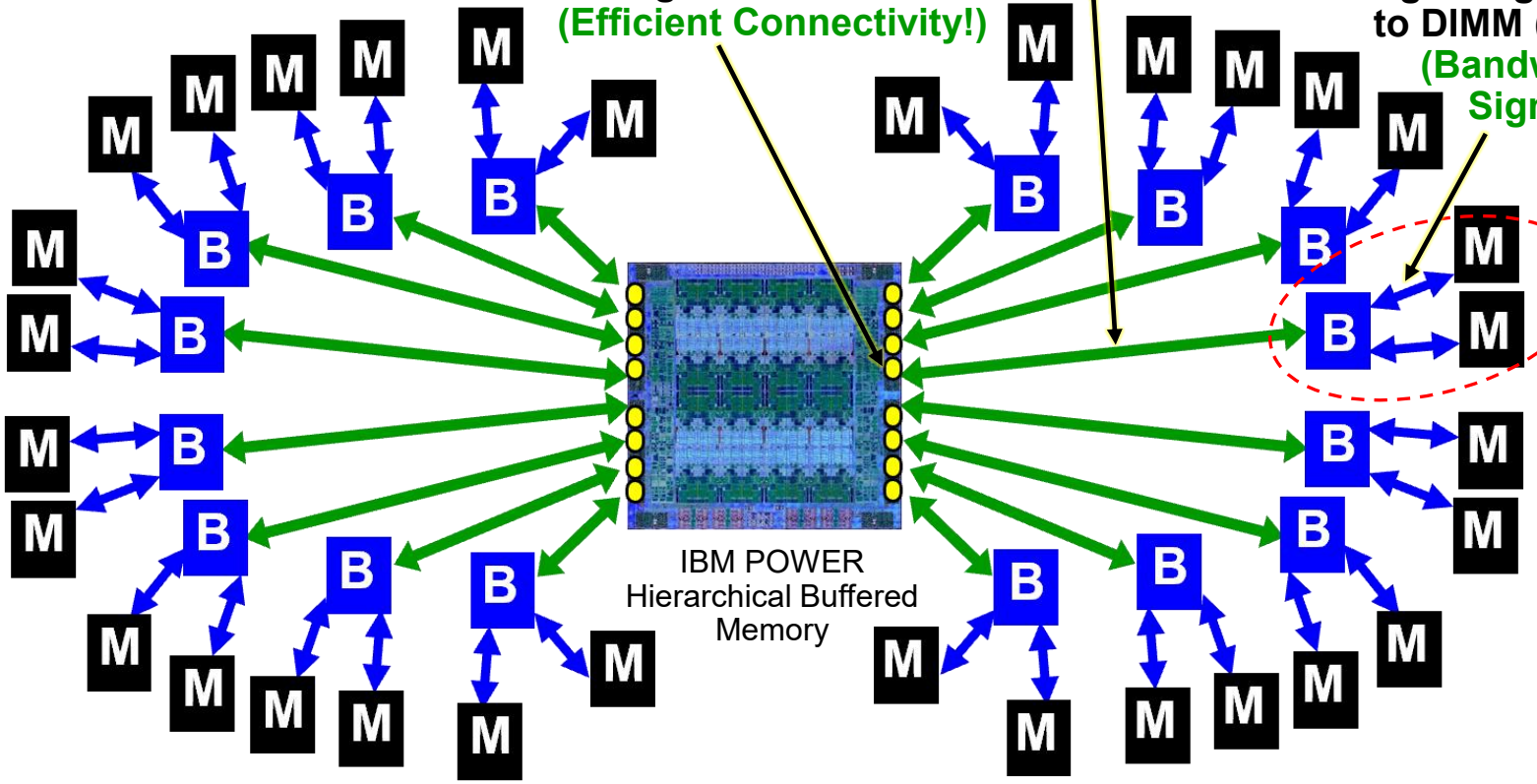
Typical Industry
Direct Attach Memory

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 single-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

Major Eras in Computing

The Clock Speed Era

The Multi-core Era

Phases of Processor / System History

The Future

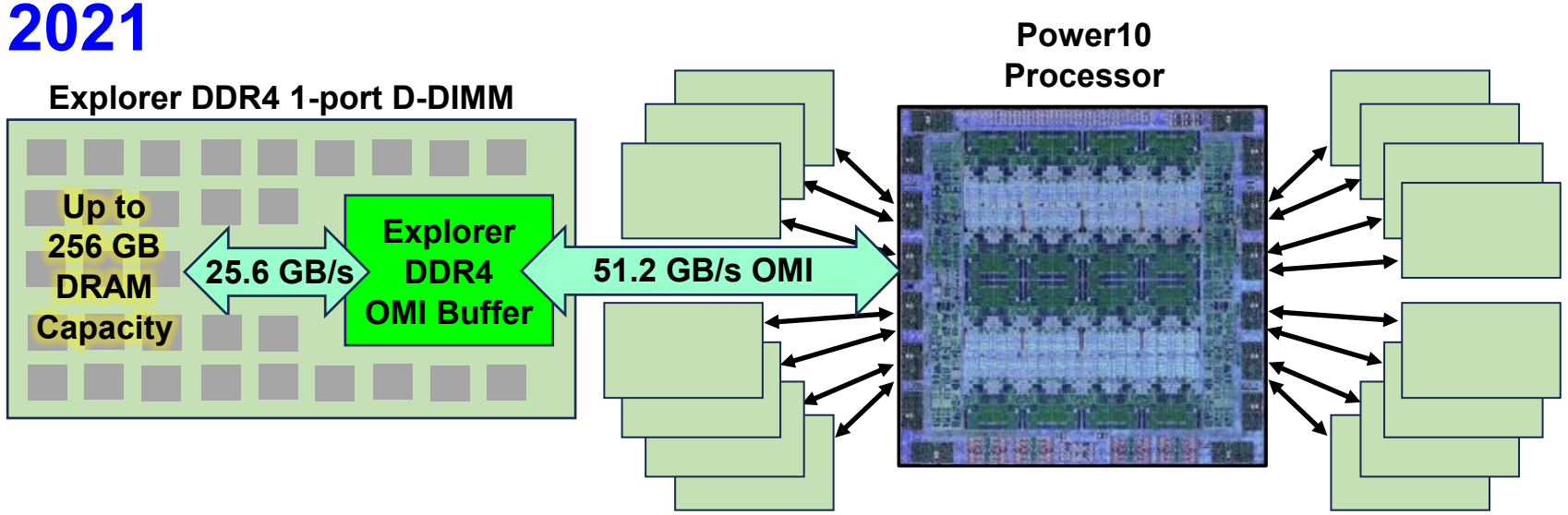
The Acceleration Era?

The Memory Era?

How will the Classic Von-Neumann bottleneck Manifest in the AI Era?

From Power10 to Power11… Strength of OMI Memory Architecture

2021

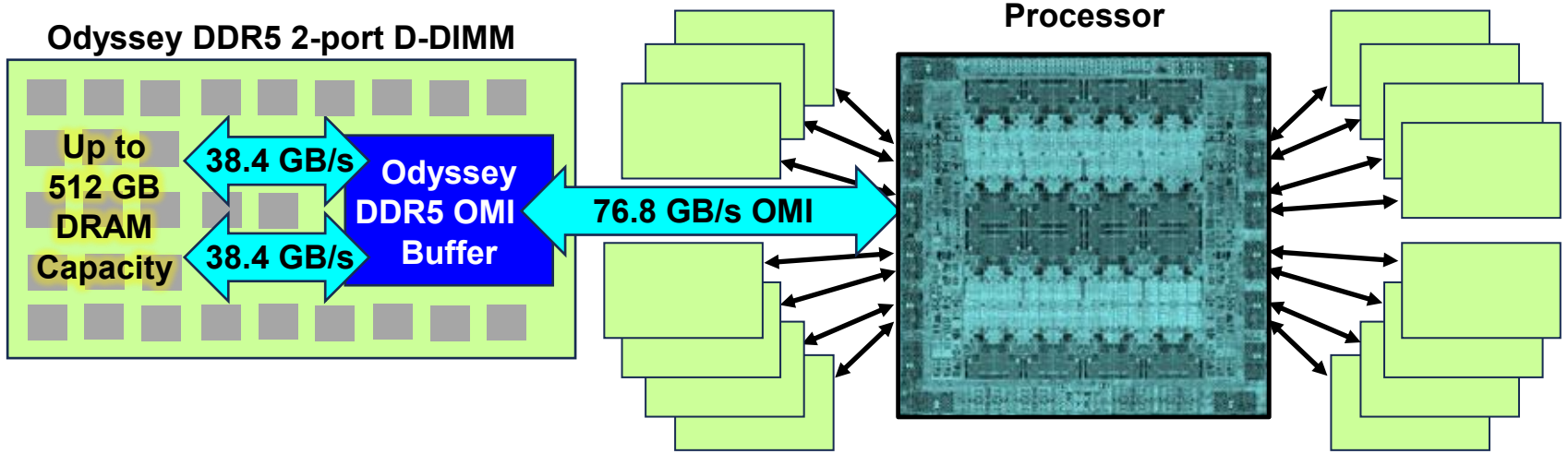


400 GB/s DRAM Bandwidth / Socket

4 Terabyte DRAM Capacity / Socket

700-800 GB/s System Coherence Flow

2025



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

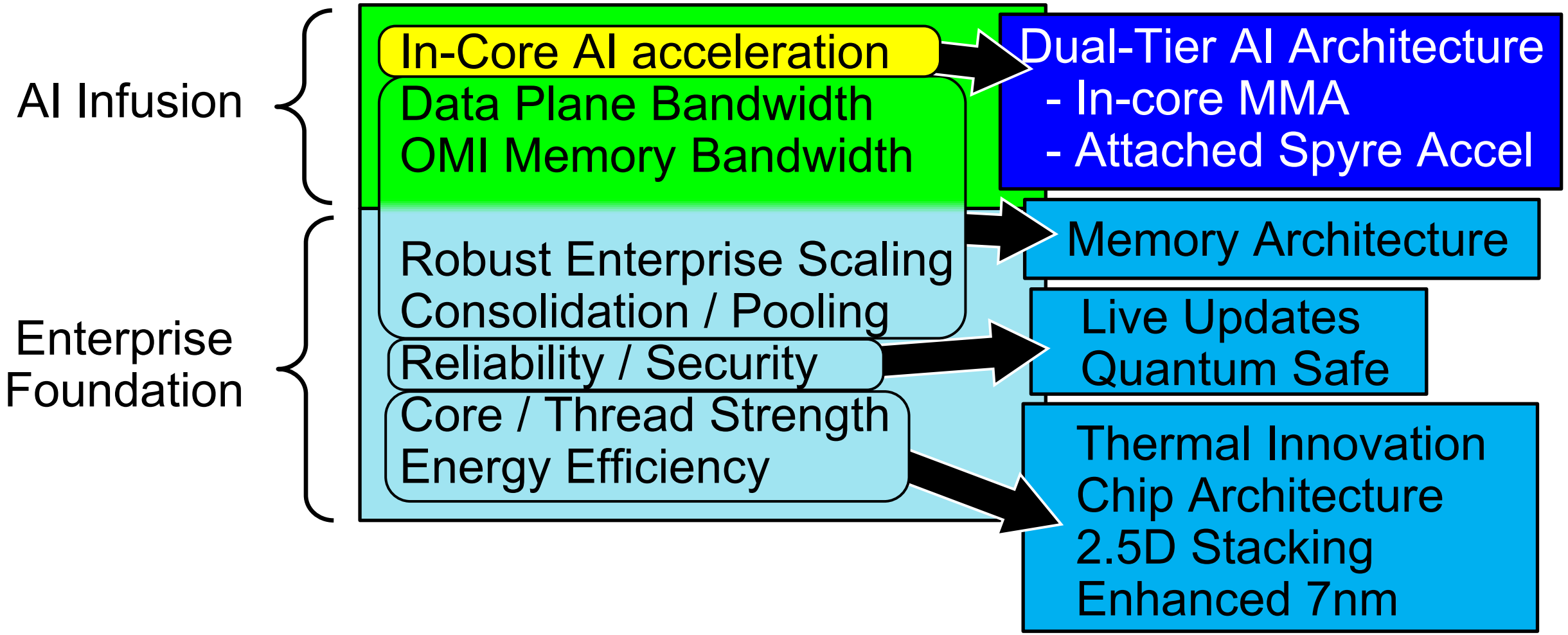
8 Terabyte DRAM Capacity / Socket **2x!!**

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

From Power10 to Power11…

Power10 Architecture:
- Tremendously Successful
- Ahead of its time for AI Infusion

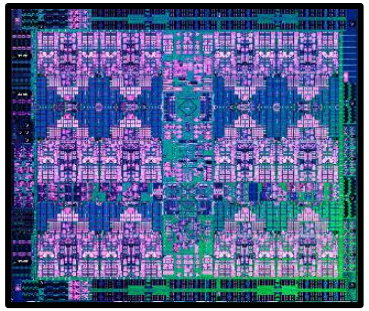
Going to 11



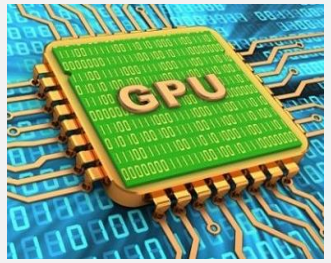
From Power10 to Power11… Accelerated AI: Large Models & Tuning

Inference | Large Inference | Fine Tuning | Training

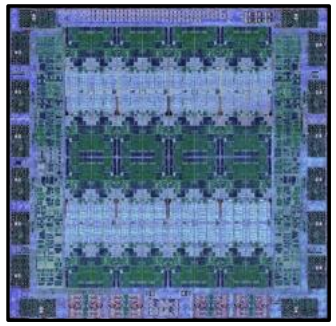
POWER9



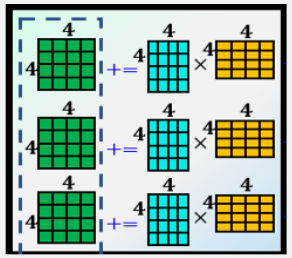
External



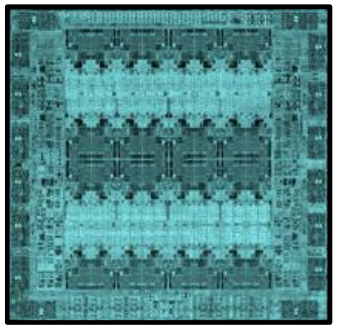
Power10



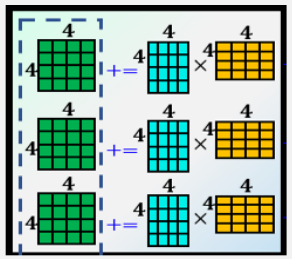
In-Core MMA



Power11



In-Core MMA



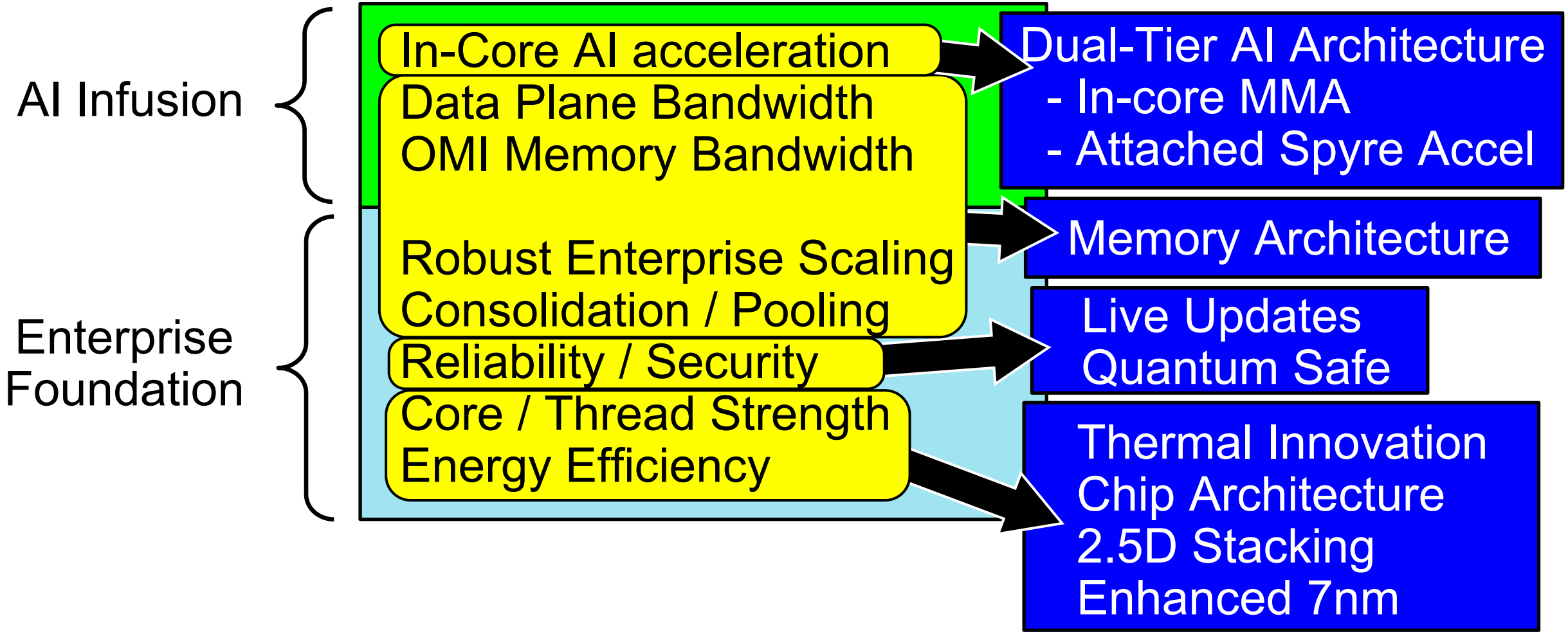
External Spyre Accelerator

From Power10 to Power11…

Power10 Architecture:
- Tremendously Successful
- Ahead of its time for AI Infusion

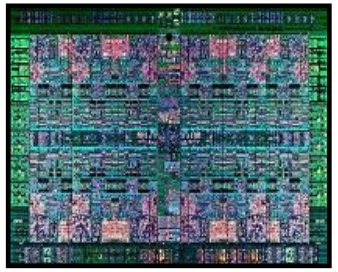
Power11 Architecture:
- Strong Foundation
- Multi-faceted Extend
- System of Hot Chip(s):

Power11, Odyssey, Spyre



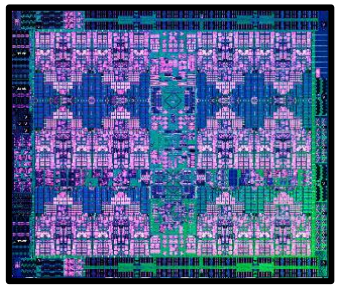
IBM Power Processor Roadmap – Beyond Power11

POWER8



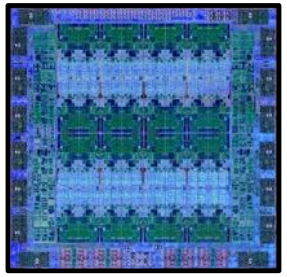
Powerful SMT8 Core
Enterprise Scaling
Big Data Optimized
Agnostic Memory

POWER9



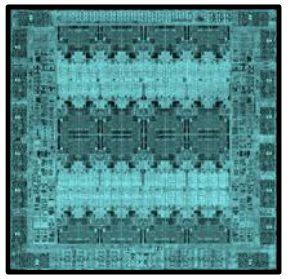
Modular Core Design
Accelerator Attach
(NVlink, OpenCAPI)
Data Plane Bandwidth
DDR & CDIMM Memory

Power10



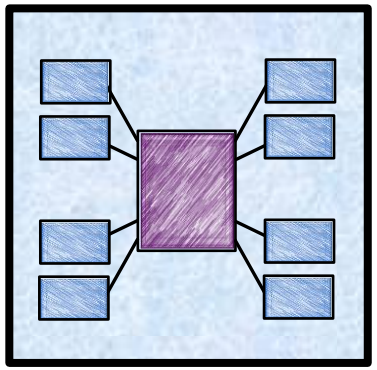
Core / Thread Strength
Socket Performance
In-Core AI Acceleration
Efficiency / Sustainability
HW Accelerated Security

Power11



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

Power Future

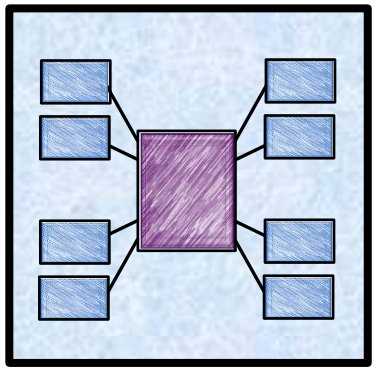


(Under development)

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

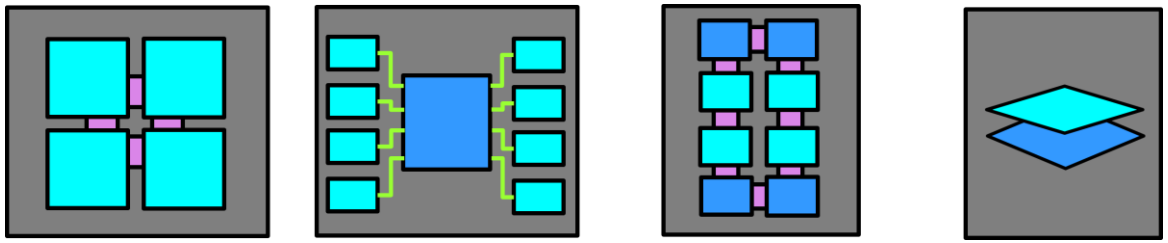
IBM Power Processor Roadmap – Beyond Power11

Power Future



(Under development)

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



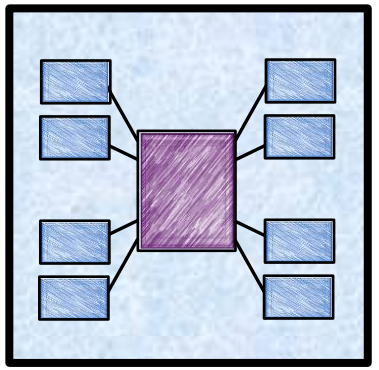
- **Several years back we assessed for Power …**

System Focus vs Socket Focus

- Full System (Socket-escape) Bandwidth vs Intra-socket Bandwidth
- Distance Signaling vs Low-power (Edge-to-edge) Signaling
- End-to-end Latency vs Computational Density
- Coherent Sharing & Pooled Capacity vs Clustering

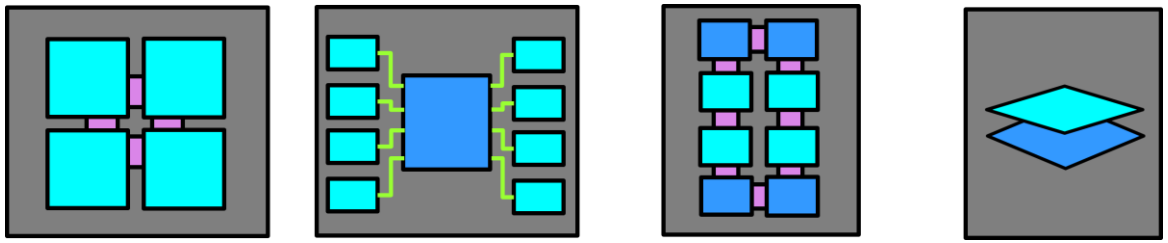
IBM Power Processor Roadmap – Beyond Power11

Power Future



(Under development)

- **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