



# POWER7 Processors: The Beat Goes On

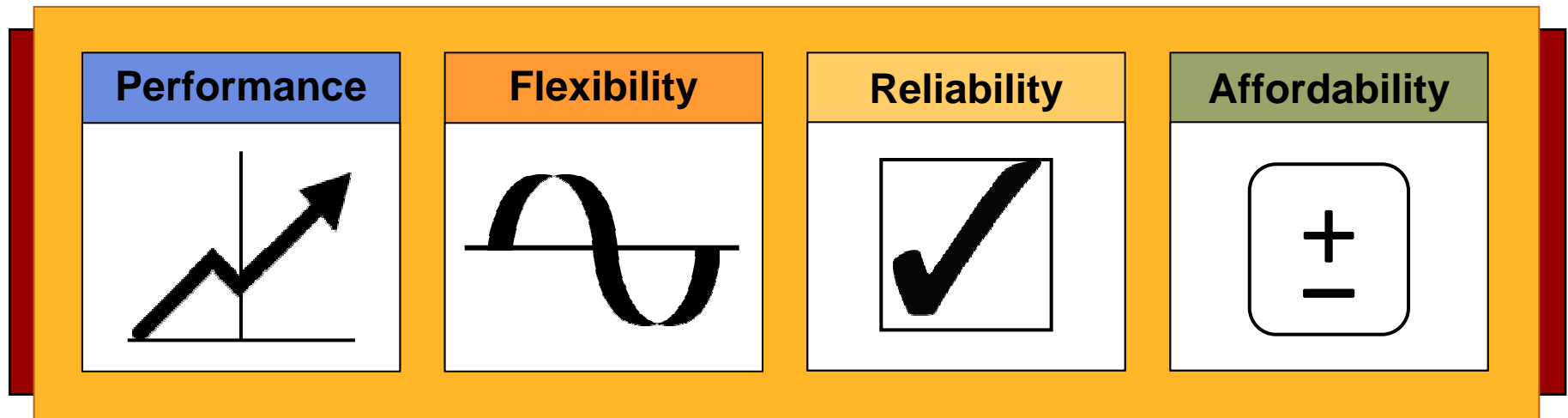
Joel M. Tandler, Executive IT Architect  
[jtandler@us.ibm.com](mailto:jtandler@us.ibm.com)

Acknowledgment: This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0002



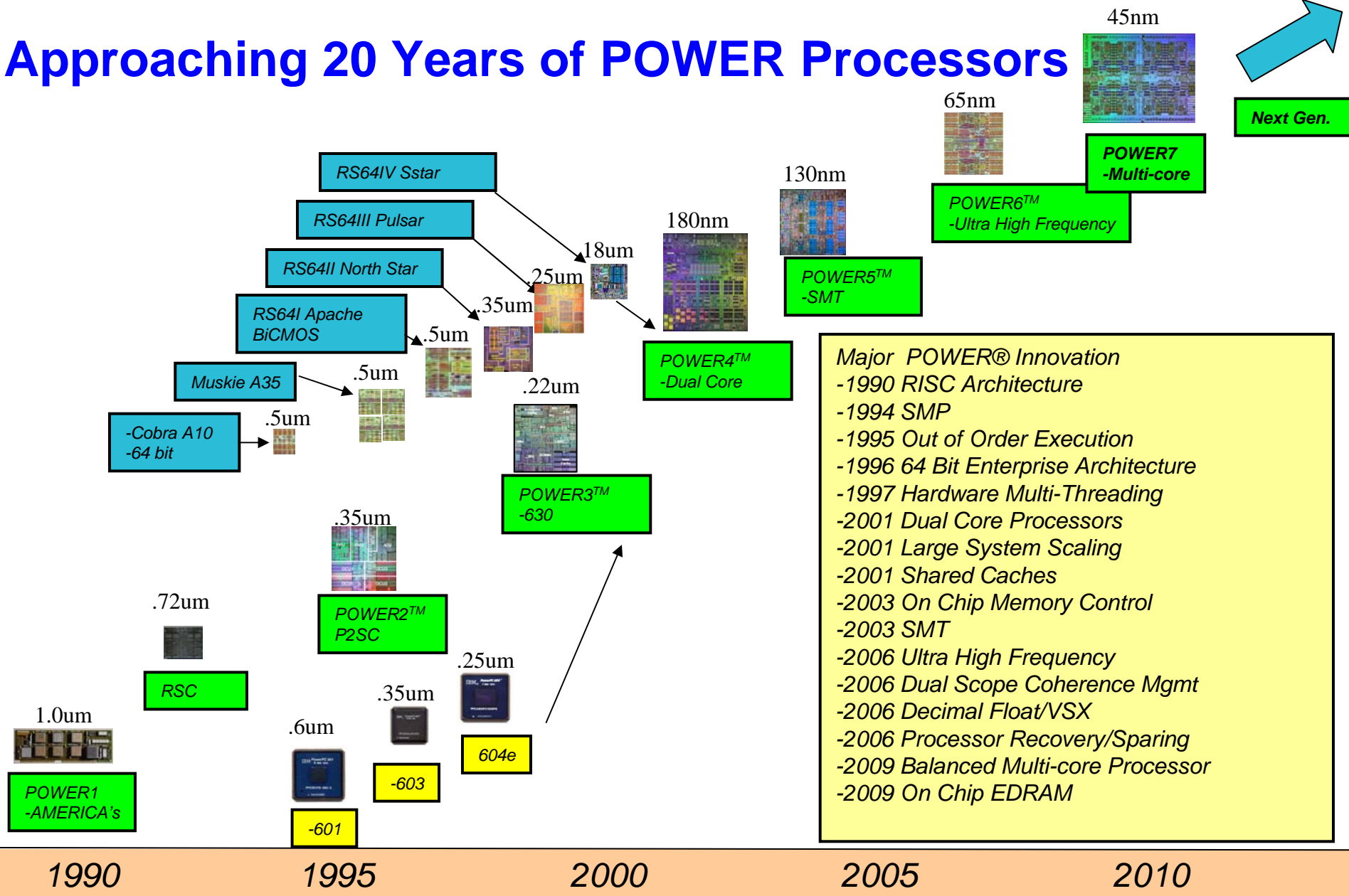
## IBM Power Systems value proposition

*Deliver business value by leveraging technology*

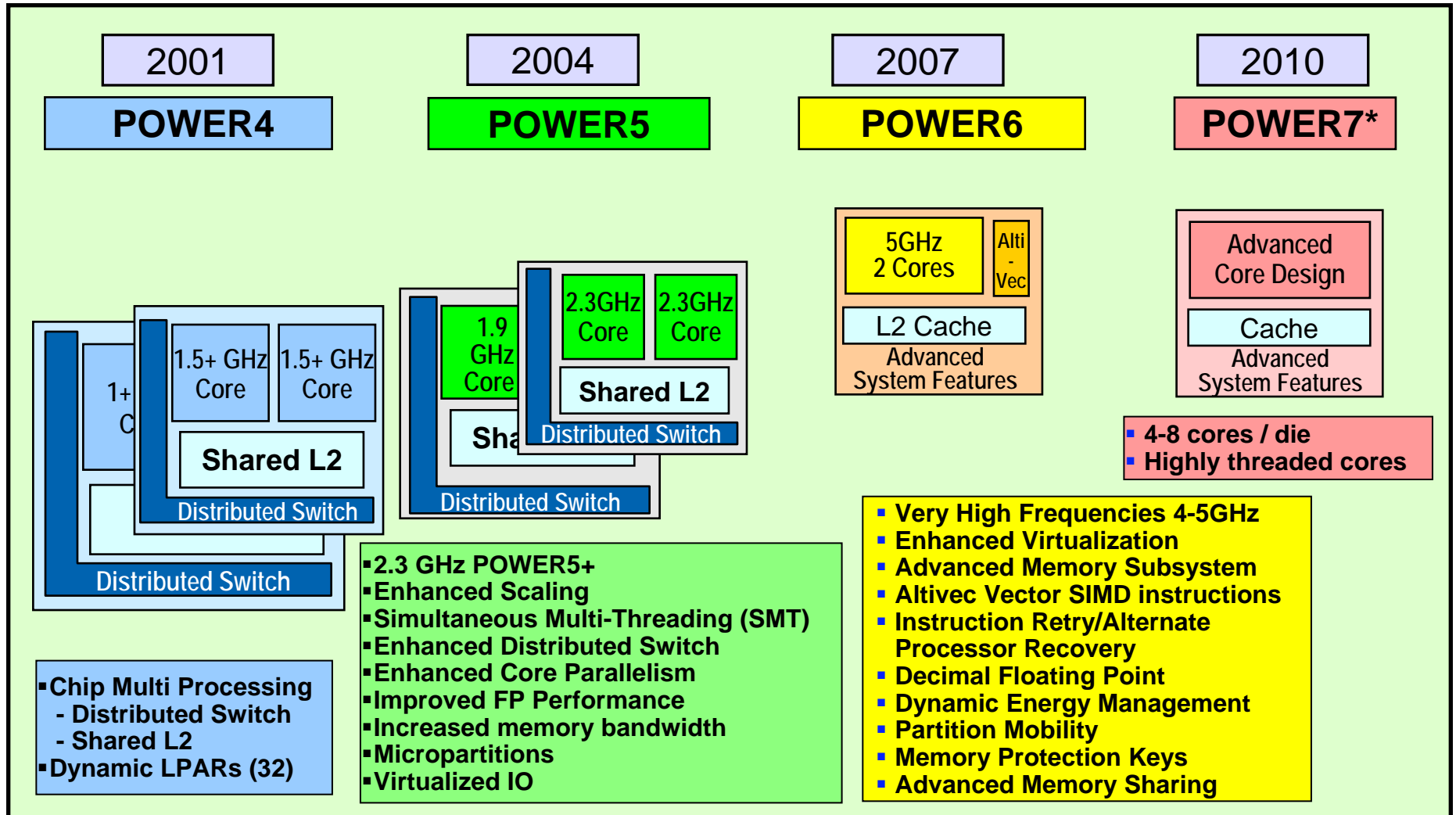


*. . . the highest value at the lowest risk  
with leading technology*

# Approaching 20 Years of POWER Processors



# POWER Roadmap – The Only Reliable Server Roadmap



First Dual core chip in industry

First Quad core in industry

Fastest chip in industry

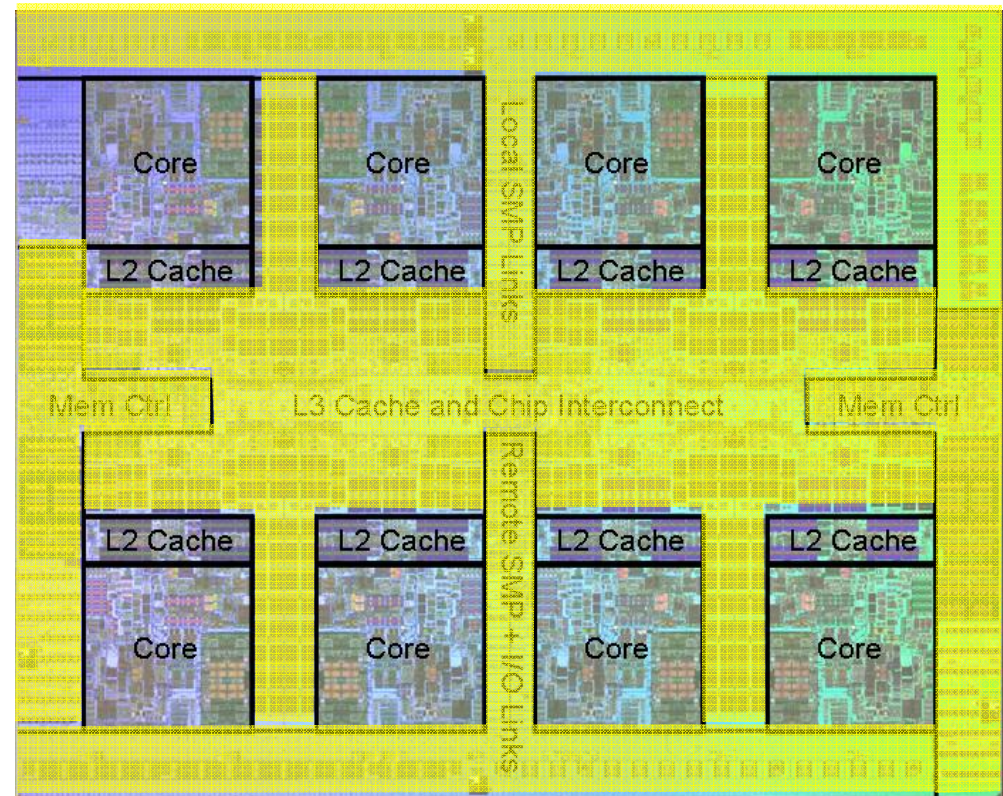
Upgrades to be available For Power 570 & Power 595



\*All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

## POWER7 Processor Chip

- 567mm<sup>2</sup> Technology: 45nm lithography, Cu, SOI, eDRAM
- 1.2B transistors
  - Equivalent function of 2.7B
  - eDRAM efficiency
- Eight processor cores
  - 12 execution units per core
  - 4 Way SMT per core
  - 32 Threads per chip
  - 256KB L2 per core
- 32MB on chip eDRAM shared L3
- Dual DDR3 Memory Controllers
  - 100GB/s Memory bandwidth per chip
- Scalability up to 32 Sockets
  - 360GB/s SMP bandwidth/chip
  - 20,000 coherent operations in flight
- Advanced pre-fetching Data and Instruction
- Binary Compatibility with POWER6 and prior systems

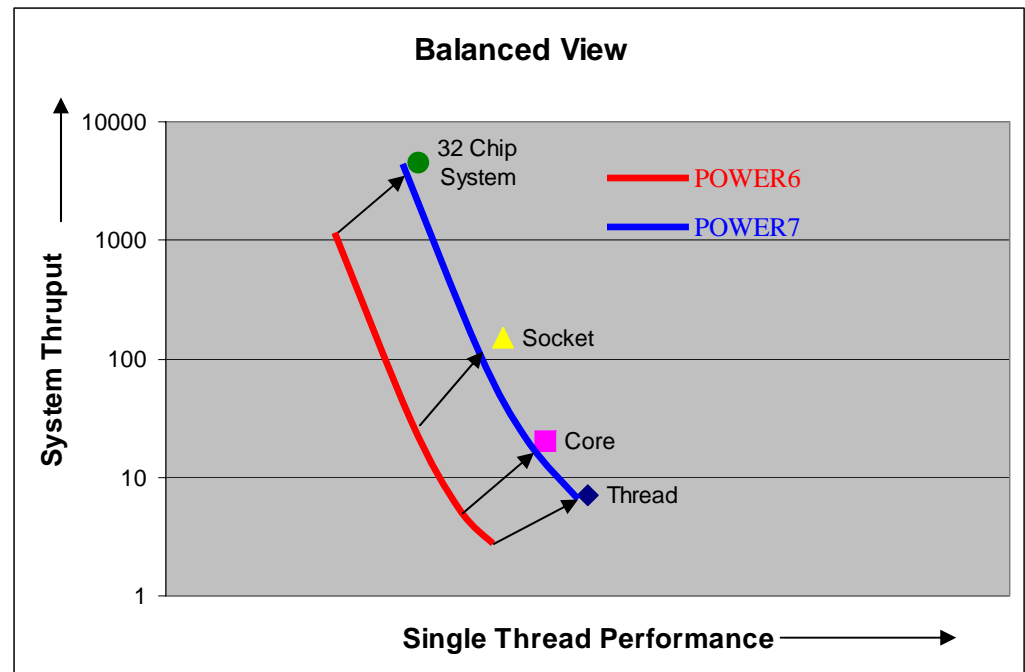
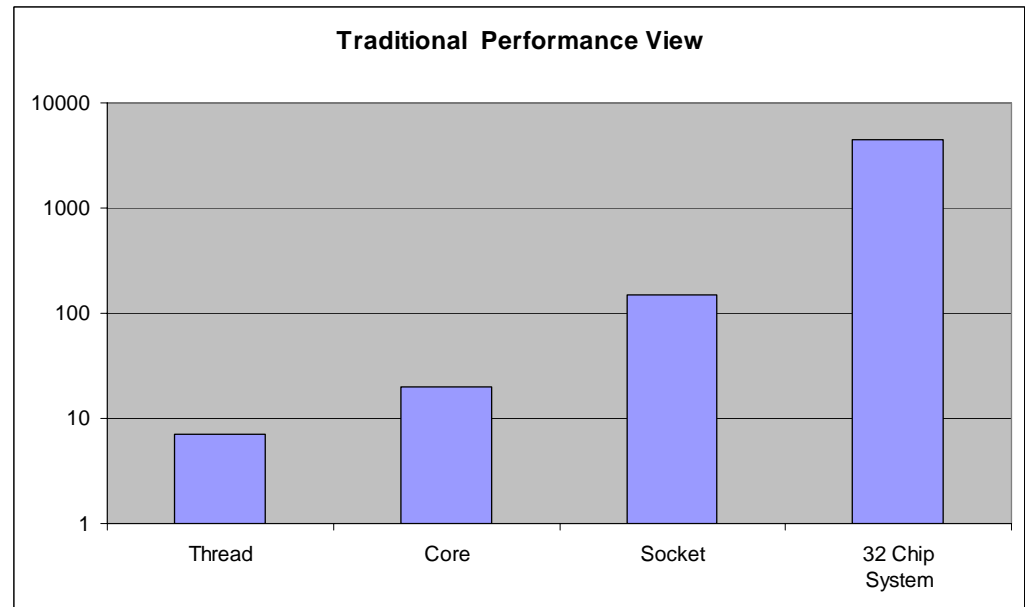


\* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.

## POWER7 Design Principles:

### Multiple optimization Points

- **Balanced Design**
  - Multiple optimization points
  - Improved energy efficiency
  - RAS improvements
- **Improved Thread Performance**
  - Dynamic allocation of resources
  - Shared L3
- **Increased Core parallelism**
  - 4 Way SMT
  - Aggressive out of order execution
- **Extreme Increase in Socket Throughput**
  - Continued growth in socket bandwidth
  - Balanced core, cache, memory improvements
- **System**
  - Scalable interconnect
  - Reduced coherence traffic



\* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.

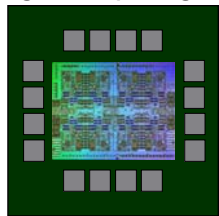
## POWER7 Design Principles:

### Flexibility and Adaptability

- Cores:
  - 8, 6, and 4-core offerings with up to 32MB of L3 Cache
  - Dynamically turn cores on and off, reallocating energy
  - Dynamically vary individual core frequencies, reallocating energy
  - Dynamically enable and disable up to 4 threads per core
- Memory Subsystem:
  - Full 8 channel or reduced 4 channel configurations
- System Topologies:
  - Standard, half-width, and double-width SMP busses supported
- Multiple System Packages

#### 2/4s Blades and Racks

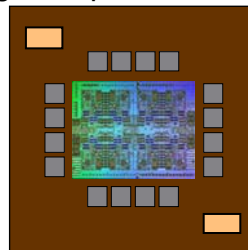
Single Chip Organic



1 Memory Controller  
3 4B local links

#### High-End and Mid-Range

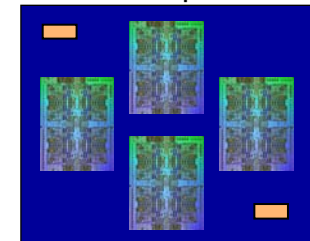
Single Chip Glass Ceramic



2 Memory Controllers  
3 8B local links  
2 8B Remote links

#### Compute Intensive

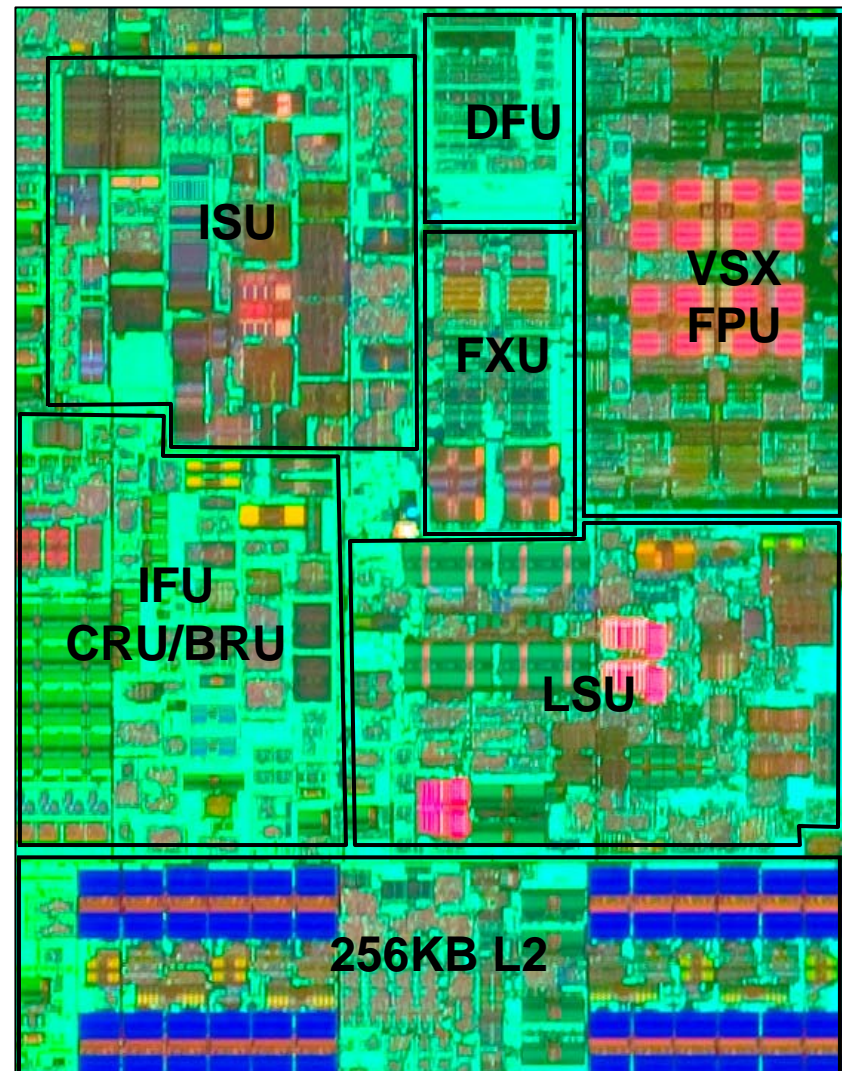
Quad-chip MCM



8 Memory Controllers  
3 16B local links (on MCM)

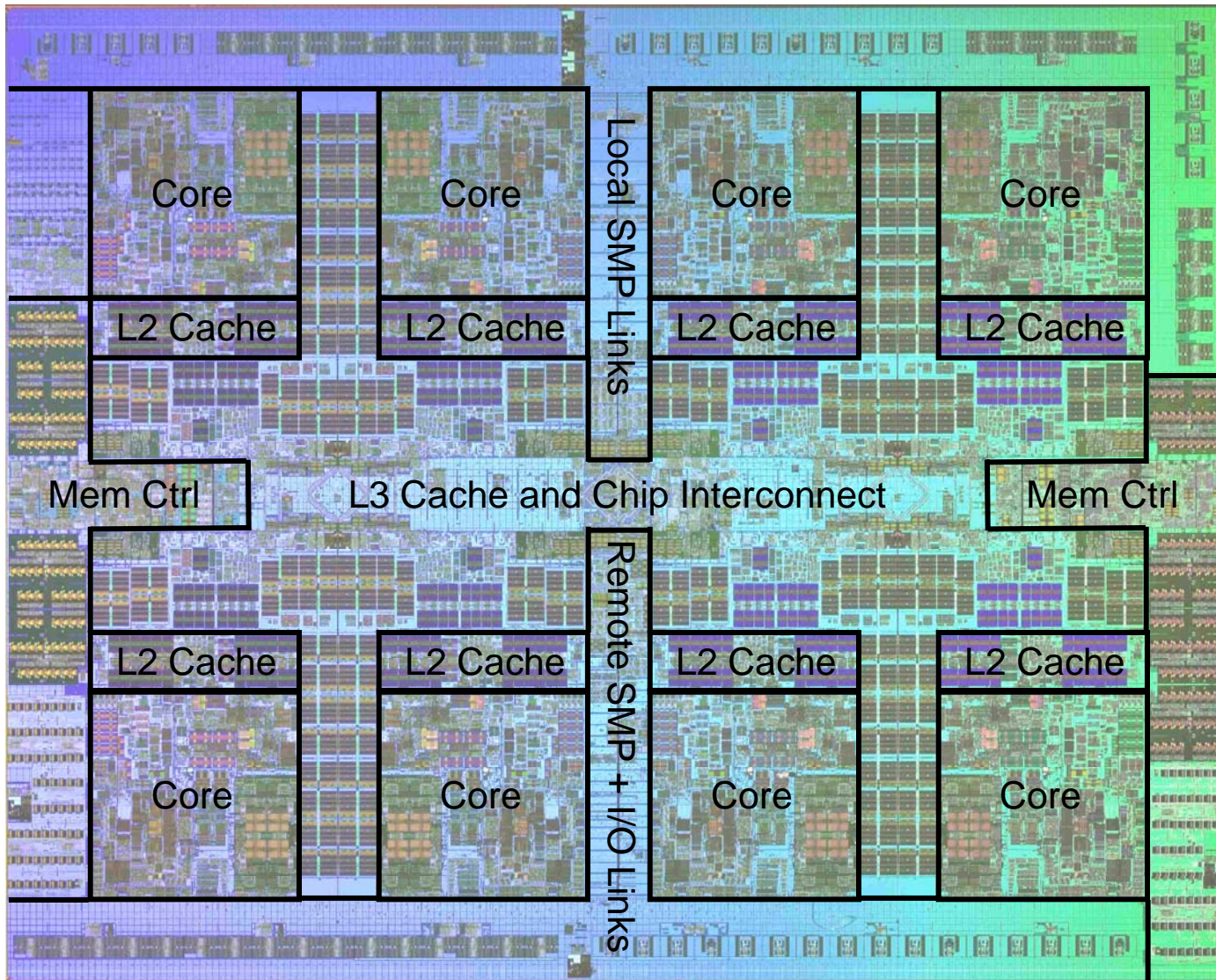
## POWER7: Core

- Execution Units
  - 2 Fixed point units
  - 2 Load store units
  - 4 Double precision floating point
  - 1 Vector unit
  - 1 Branch
  - 1 Condition register
  - 1 Decimal floating point unit
  - 6 Wide dispatch/8 Wide Issue
- Recovery Function Distributed
- 1,2,4 Way SMT Support
- Out of Order Execution
- 32KB I-Cache
- 32KB D-Cache
- 256KB L2
  - Tightly coupled to core



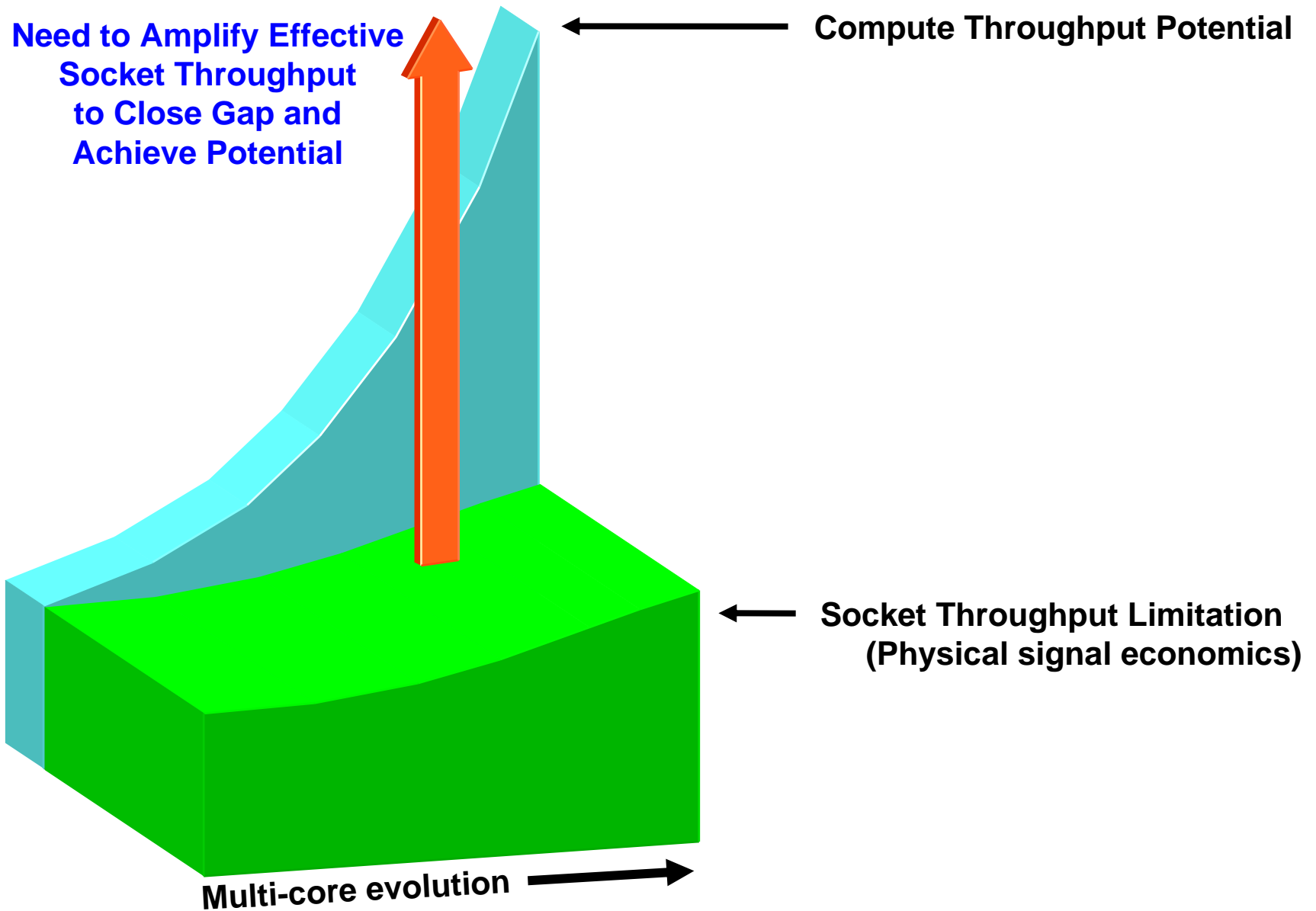


## Challenge: Beating Physics to Realize Multi-core Potential

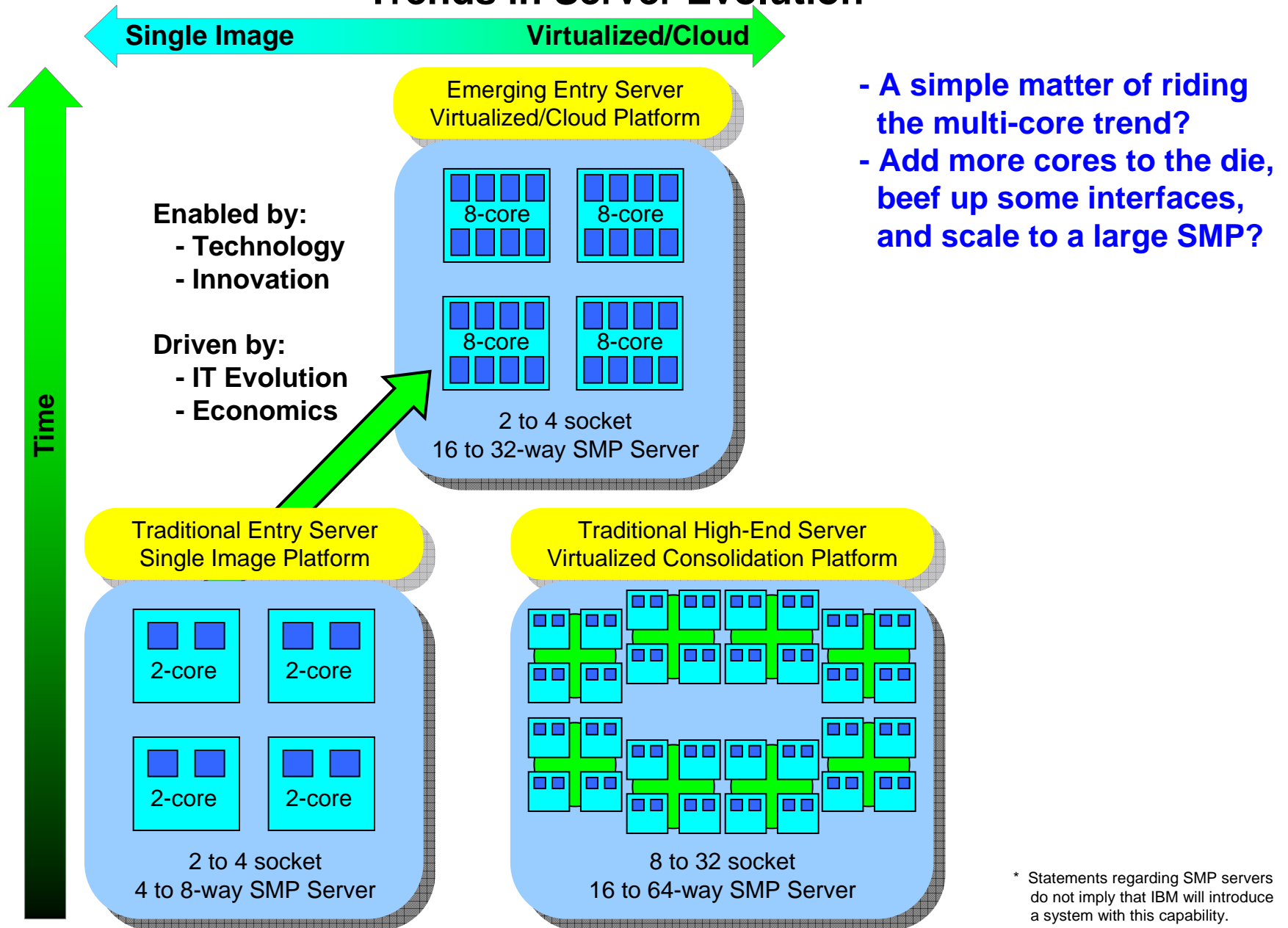


**POWER7™ is an 8-core, high performance Server chip. A solid chip is a good start. But to win the race, you need a balanced system. POWER7 enables that balance.**

## Challenge: Beating Physics to Realize Multi-core Potential

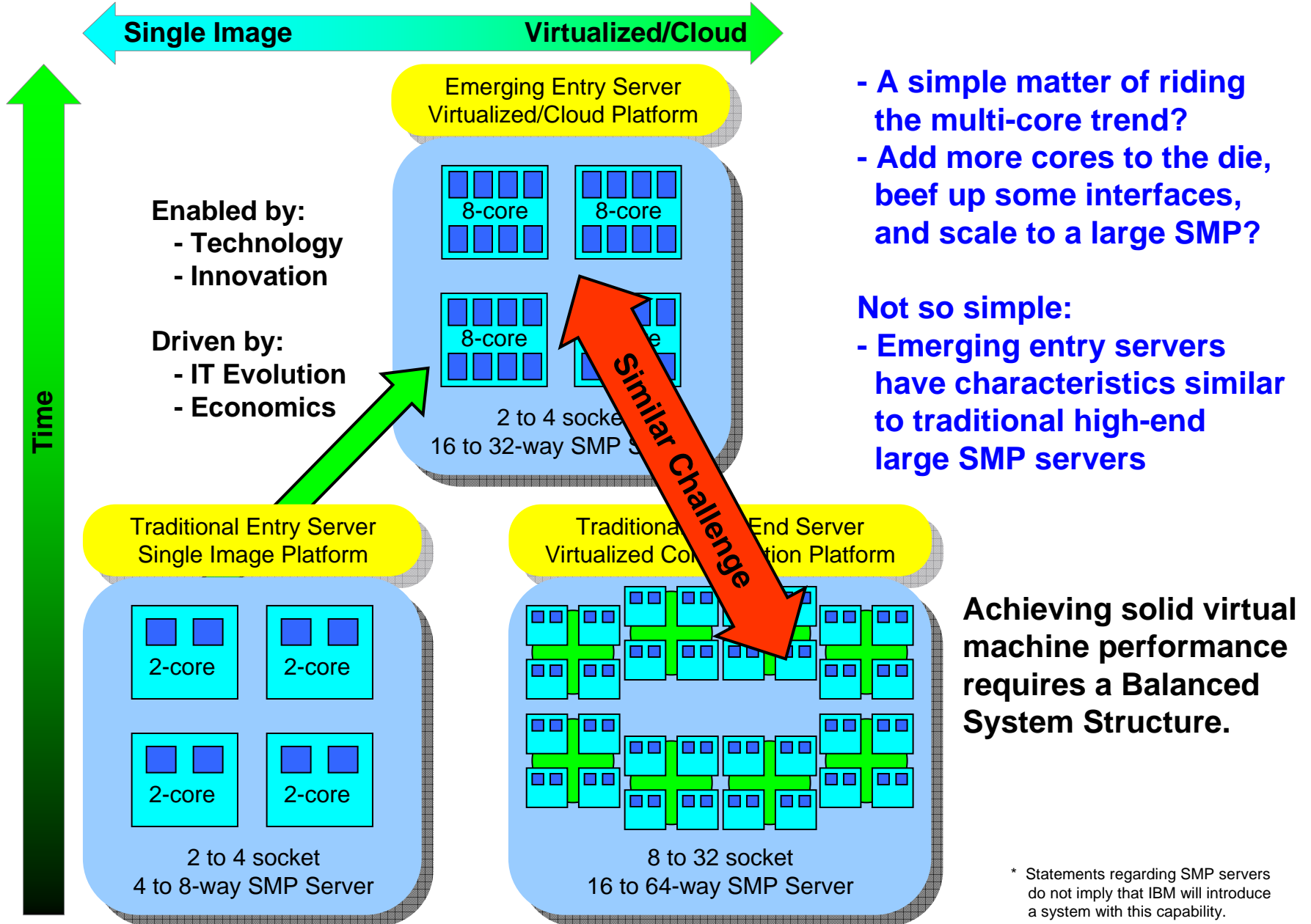


# Trends in Server Evolution

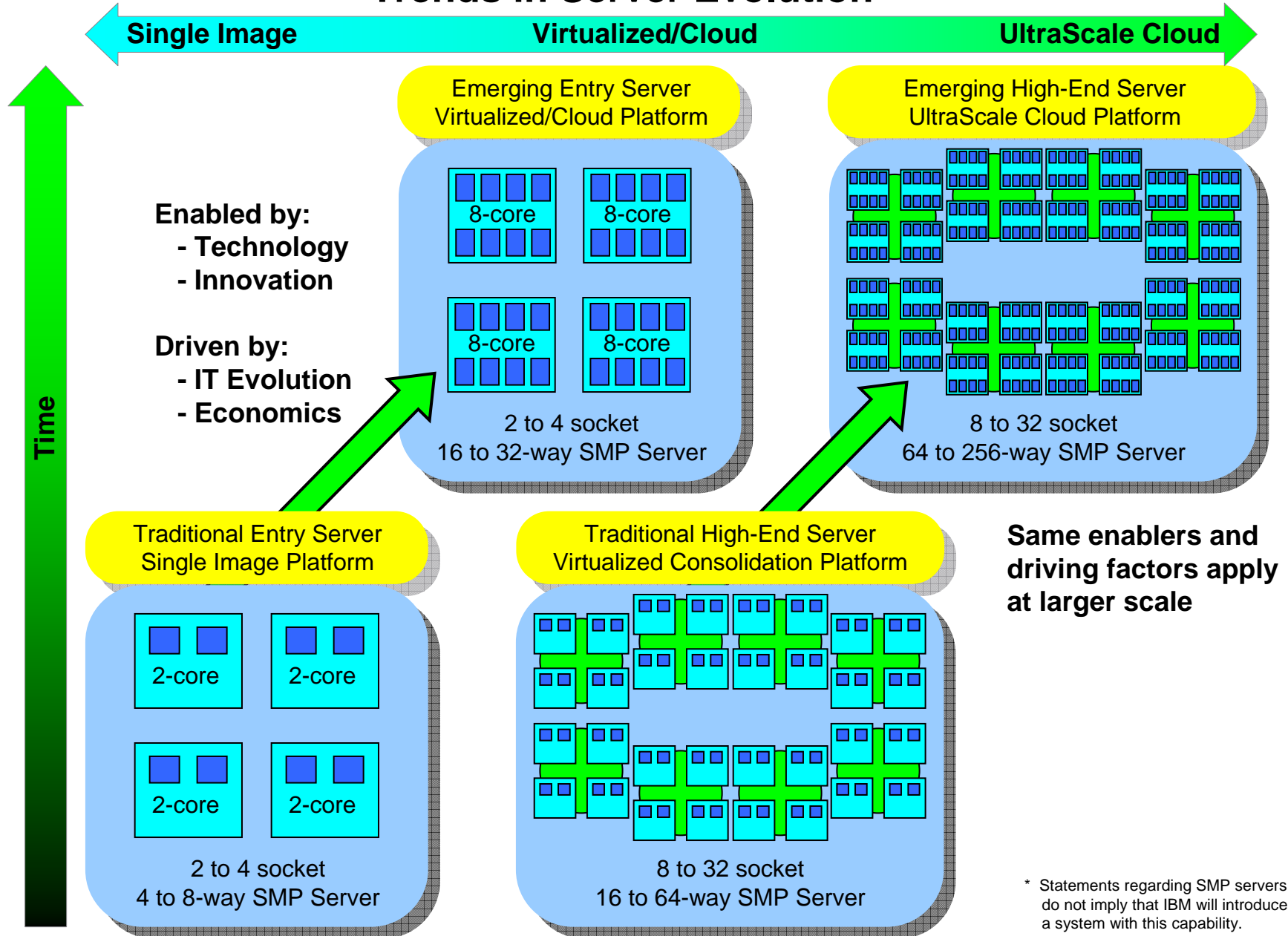


\* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.

# Trends in Server Evolution

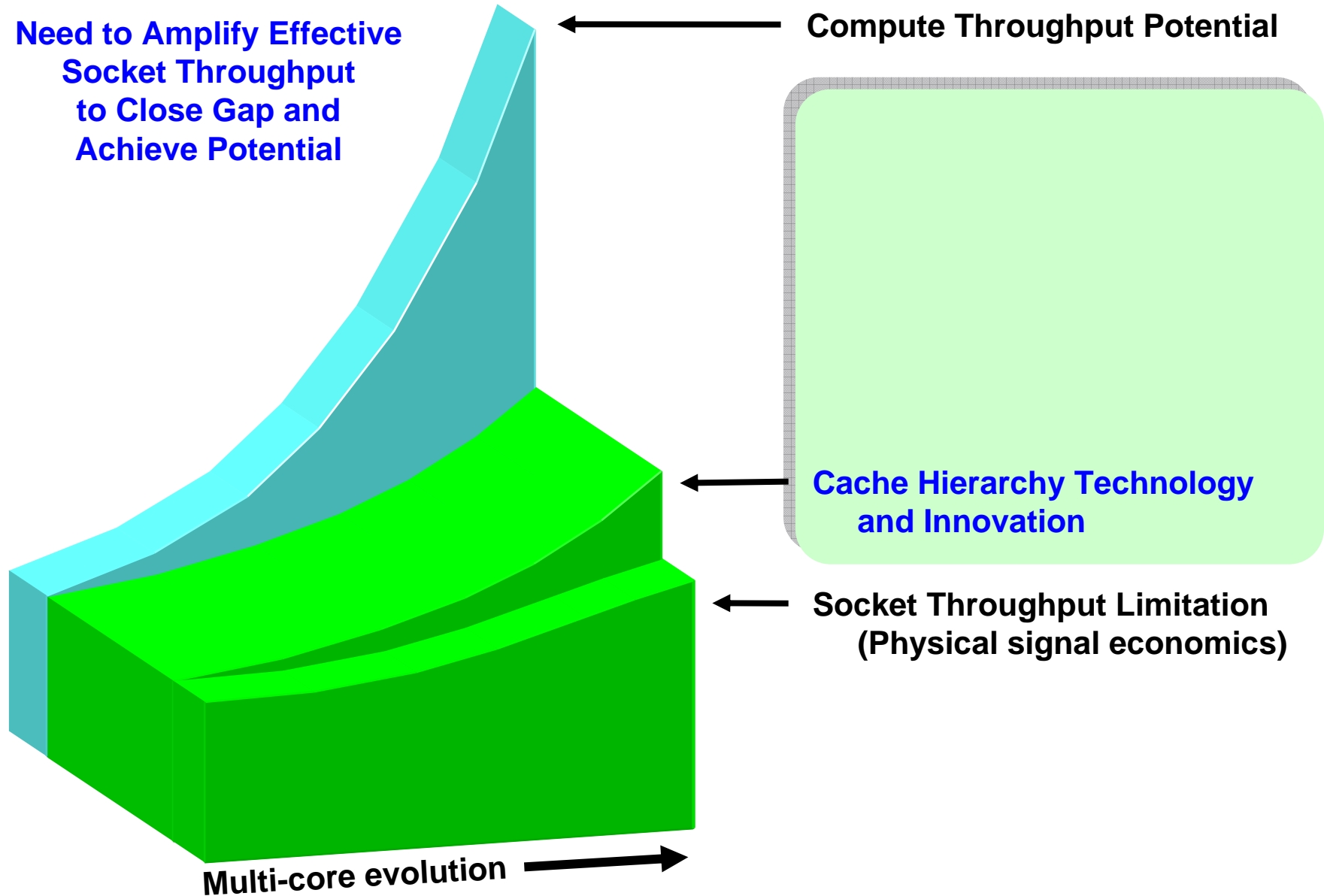


# Trends in Server Evolution



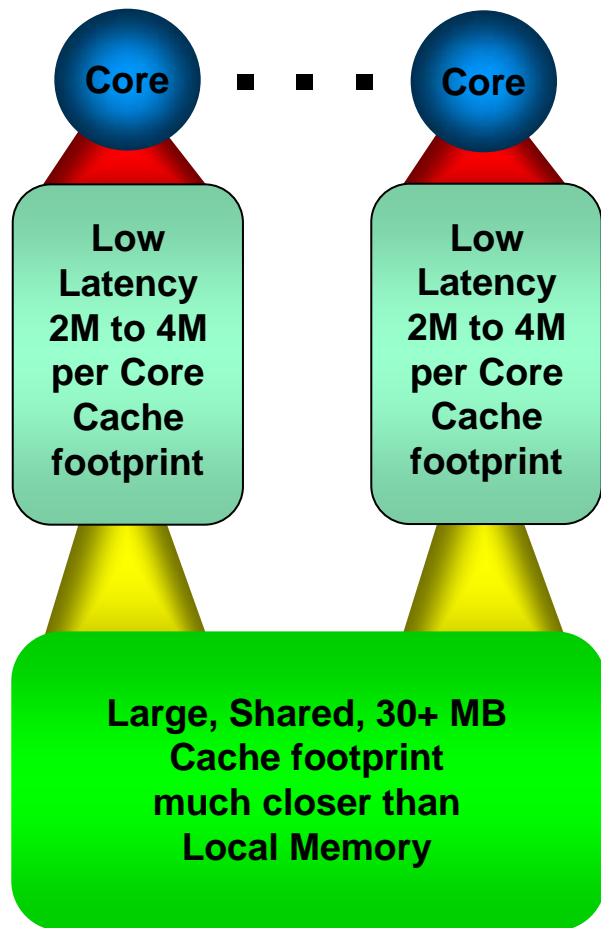
# Challenge: How does POWER7 maintain the Balance?

**Need to Amplify Effective Socket Throughput to Close Gap and Achieve Potential**



## Cache Hierarchy Technology and Innovation

### Cache Hierarchy Rqmt for POWER® Servers



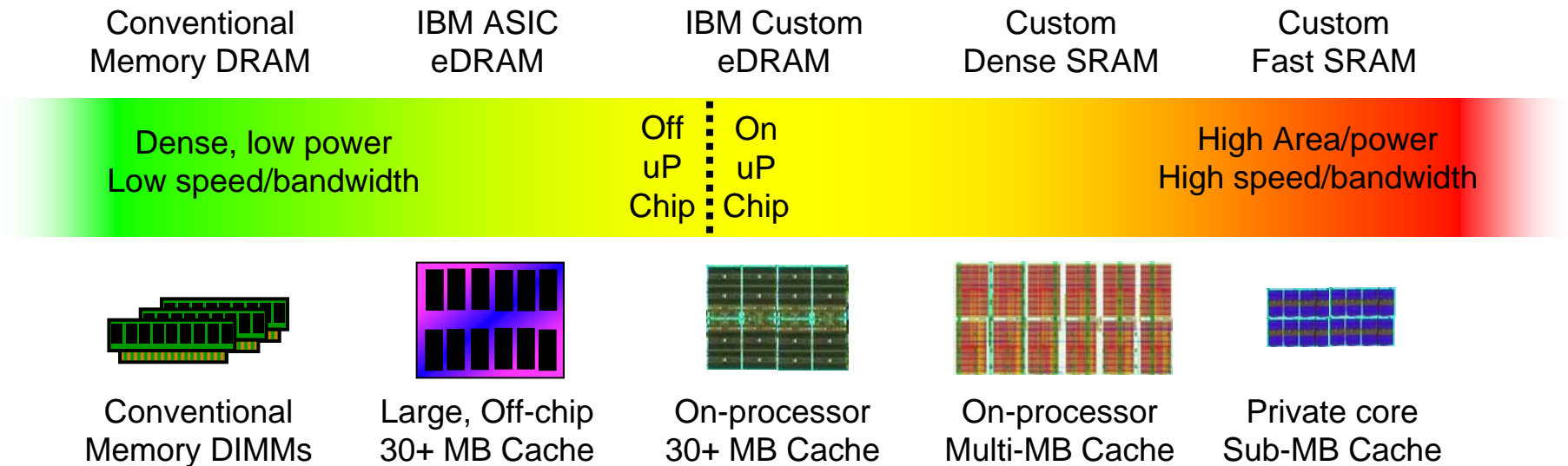
### Challenge for Multi-core POWER7

POWER4™, POWER5™, and POWER6™ systems derive huge benefit from high bandwidth access to large, off-chip cache.

But socket pin count constraints prevent scaling the off-chip cache interface to support 8 cores.

## Cache Hierarchy Technology and Innovation

**Solution: High speed eDRAM on the processor die**



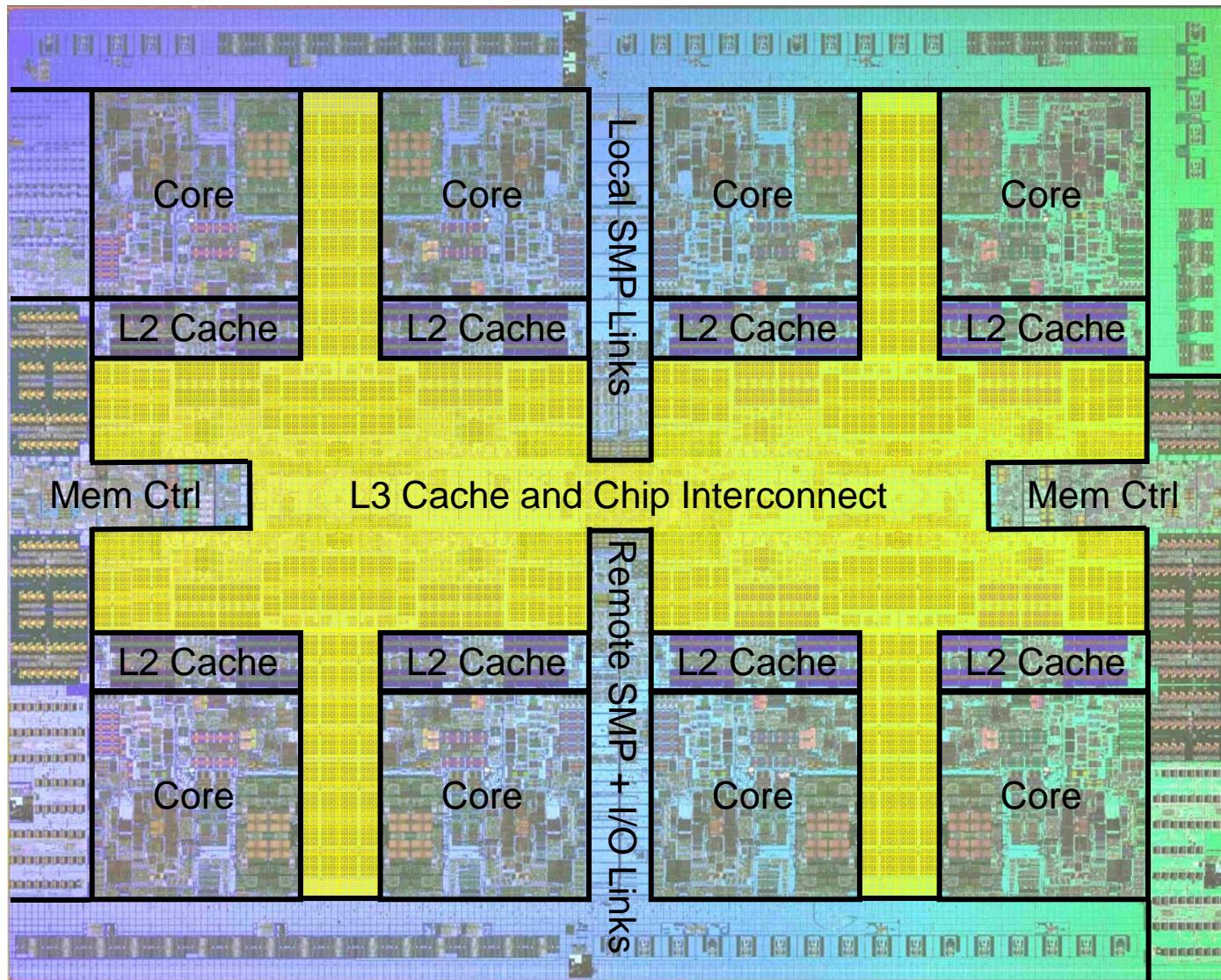
**Industry Standard Caching and Memory Technologies:  
Conventional DIMMs, Dense and Fast SRAM's.**

**IBM's POWER Servers have leveraged large off-chip eDRAM caches in POWER4, 5, and 6.**

**With POWER7, IBM introduces on-processor, high-speed, custom eDRAM, combining the dense, low power attributes of eDRAM with the speed and bandwidth of SRAM.**



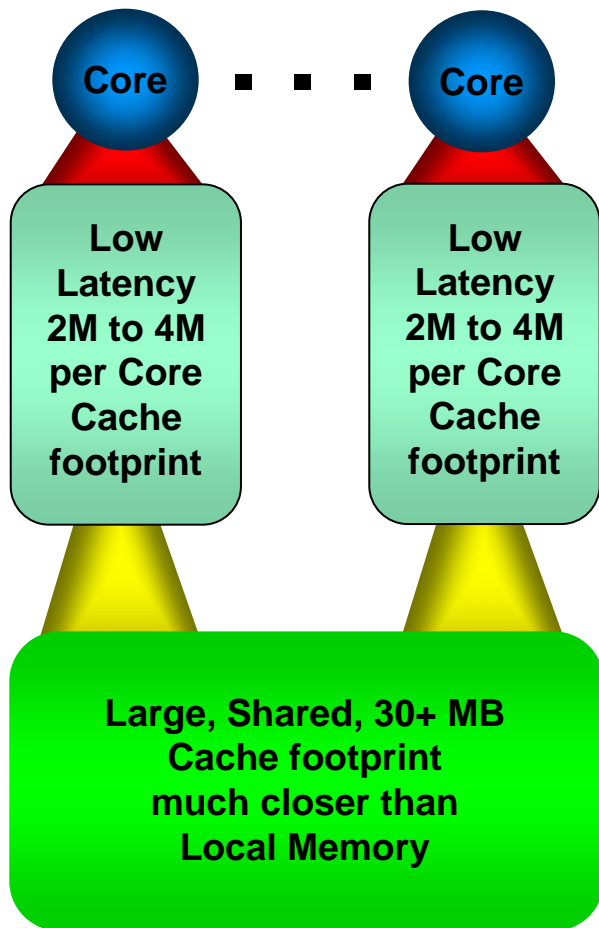
# Cache Hierarchy Technology and Innovation



## Cache Hierarchy Technology and Innovation

### Cache Hierarchy Rqmt for POWER Servers

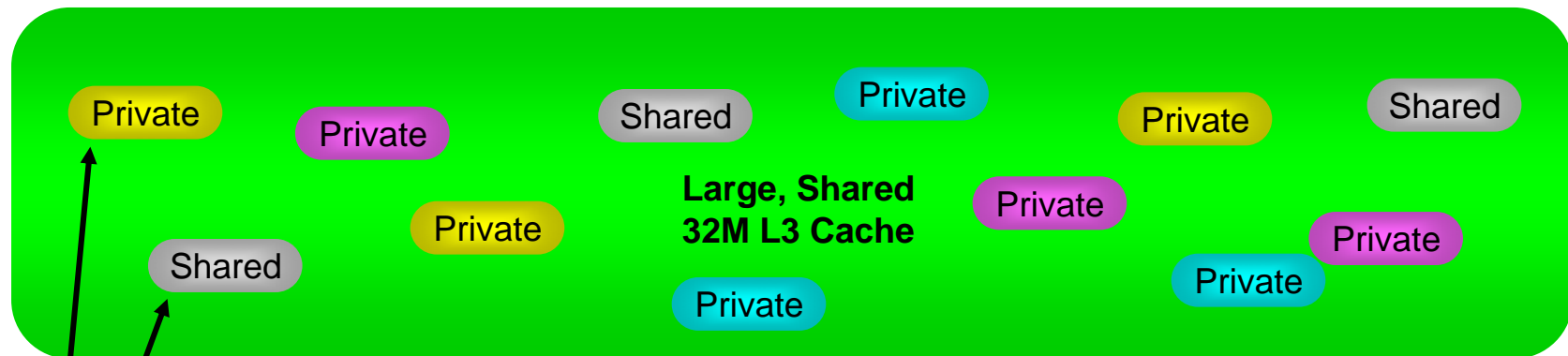
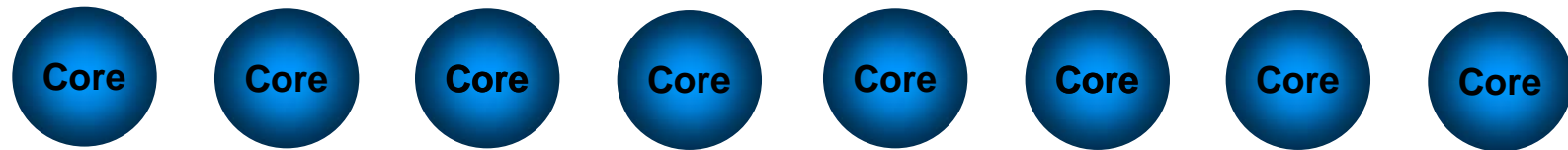
### Challenge for Multi-core POWER7



Need to satisfy both caching requirements with one cache.

# Cache Hierarchy Technology and Innovation

## Solution: Hybrid L3 "Fluid" Cache Structure

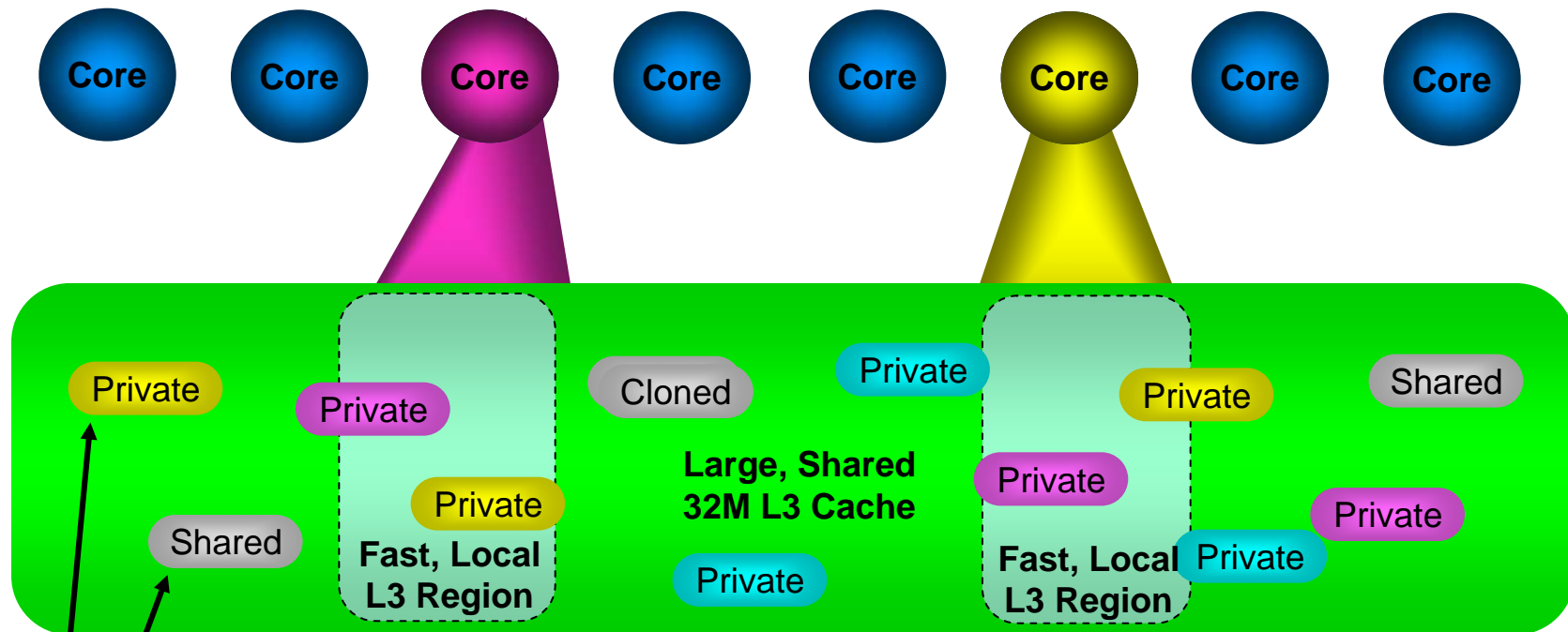


- Keeps multiple footprints at ~3X lower latency than local memory.

Working Set  
Footprints

# Cache Hierarchy Technology and Innovation

## Solution: Hybrid L3 “Fluid” Cache Structure

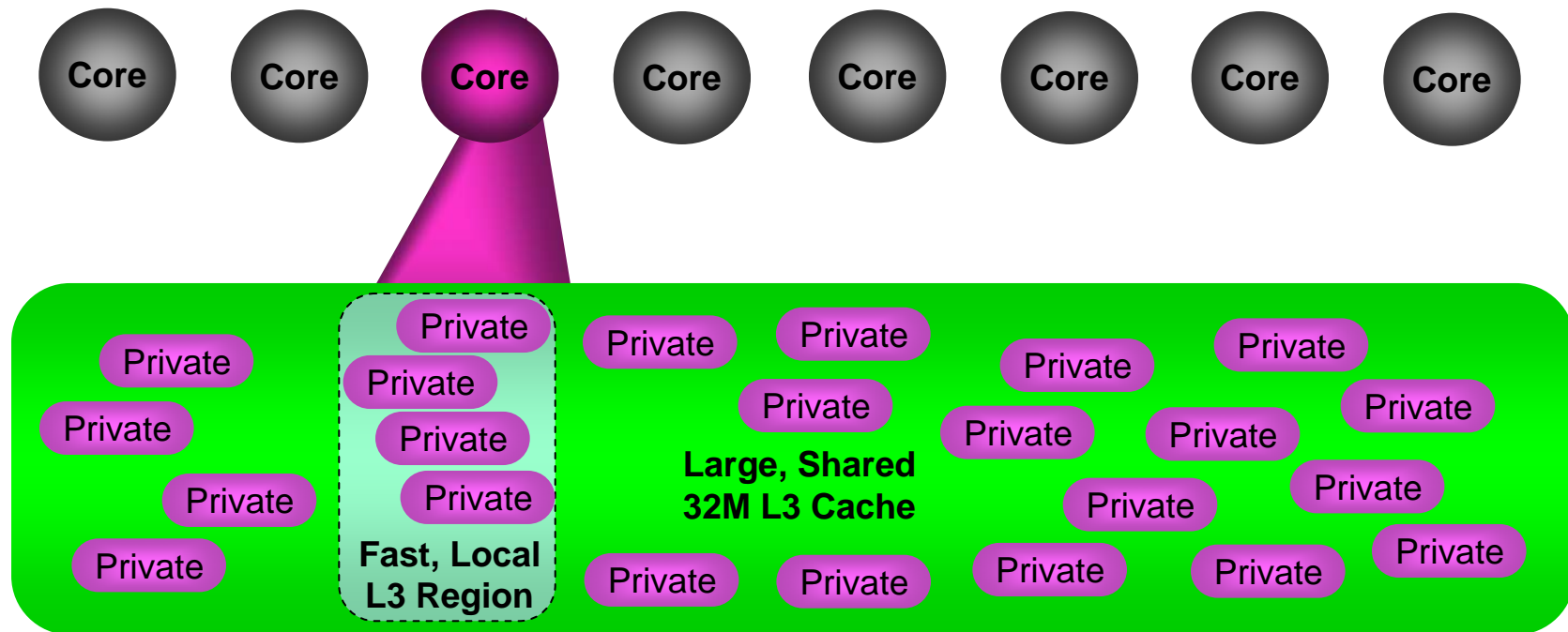


Working Set Footprints

- Keeps multiple footprints at ~3X lower latency than local memory.
- Automatically migrates private footprints (up to 4M) to fast local region (per core) at ~5X lower latency than full L3 cache.
- Automatically clones shared data to multiple private regions.

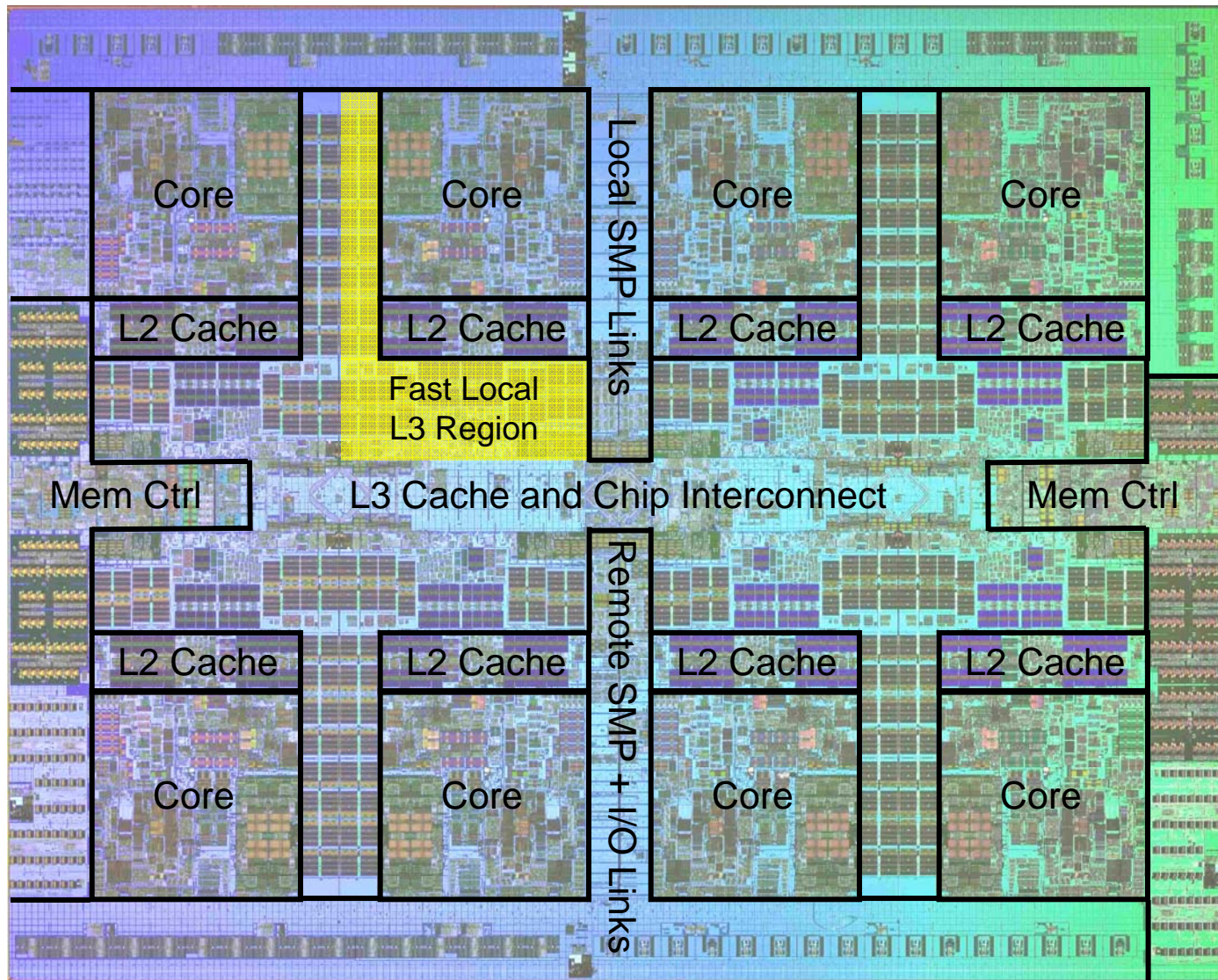
## Cache Hierarchy Technology and Innovation

### Solution: Hybrid L3 “Fluid” Cache Structure



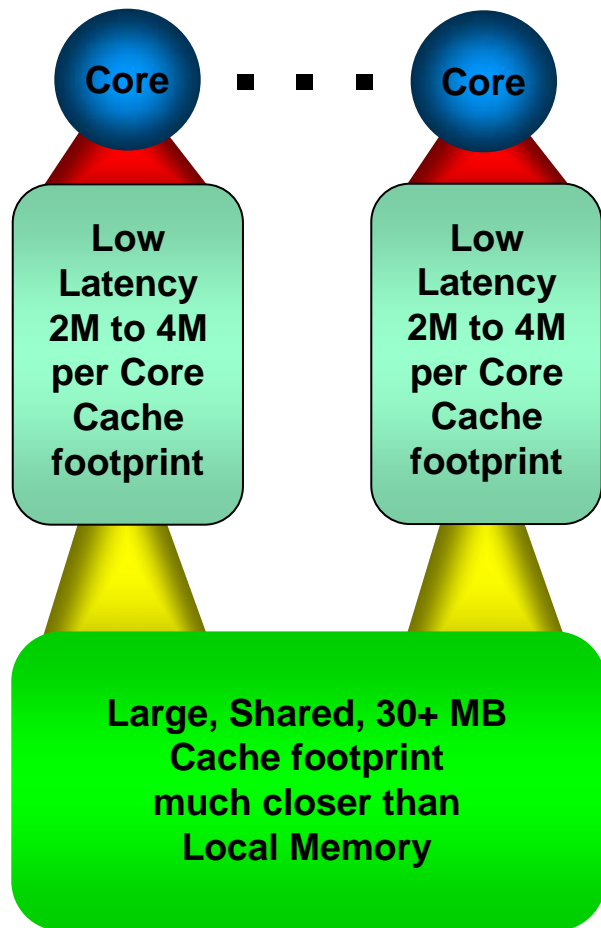
- Enables a subset of the cores to utilize the entire large shared L3 cache when the remaining cores are not using it.

## Cache Hierarchy Technology and Innovation



# Cache Hierarchy Technology and Innovation

## Cache Hierarchy Rqmt for POWER Servers



## Challenge for Multi-core POWER7

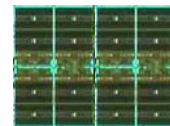
Low power, dense eDRAM value enhanced with low latency, high bandwidth, fast SRAM structures

IBM Custom eDRAM

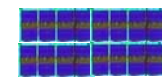
Custom Fast SRAM

Dense, low power  
Lower speed/bandwidth

High Area/power  
High speed/bandwidth



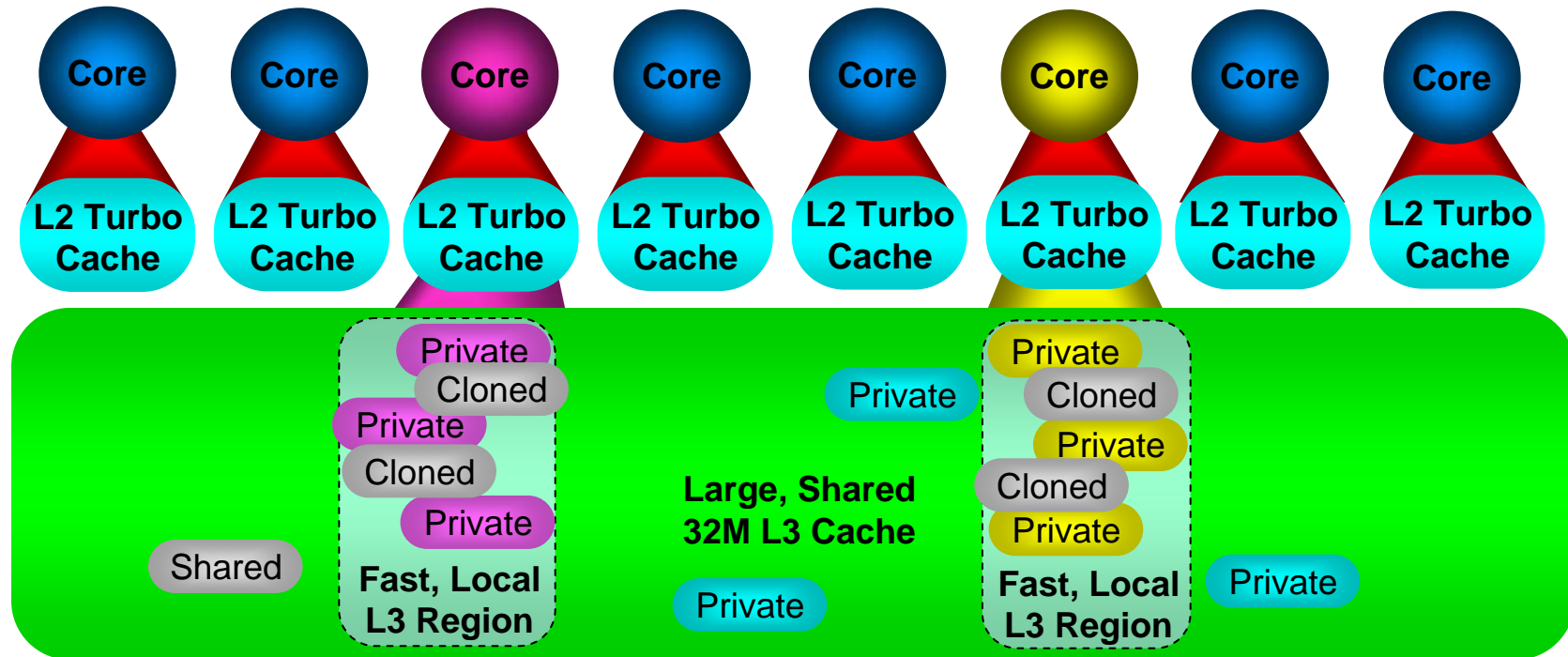
On-processor  
30+ MB Cache



Private core  
Sub-MB Cache

## Cache Hierarchy Technology and Innovation

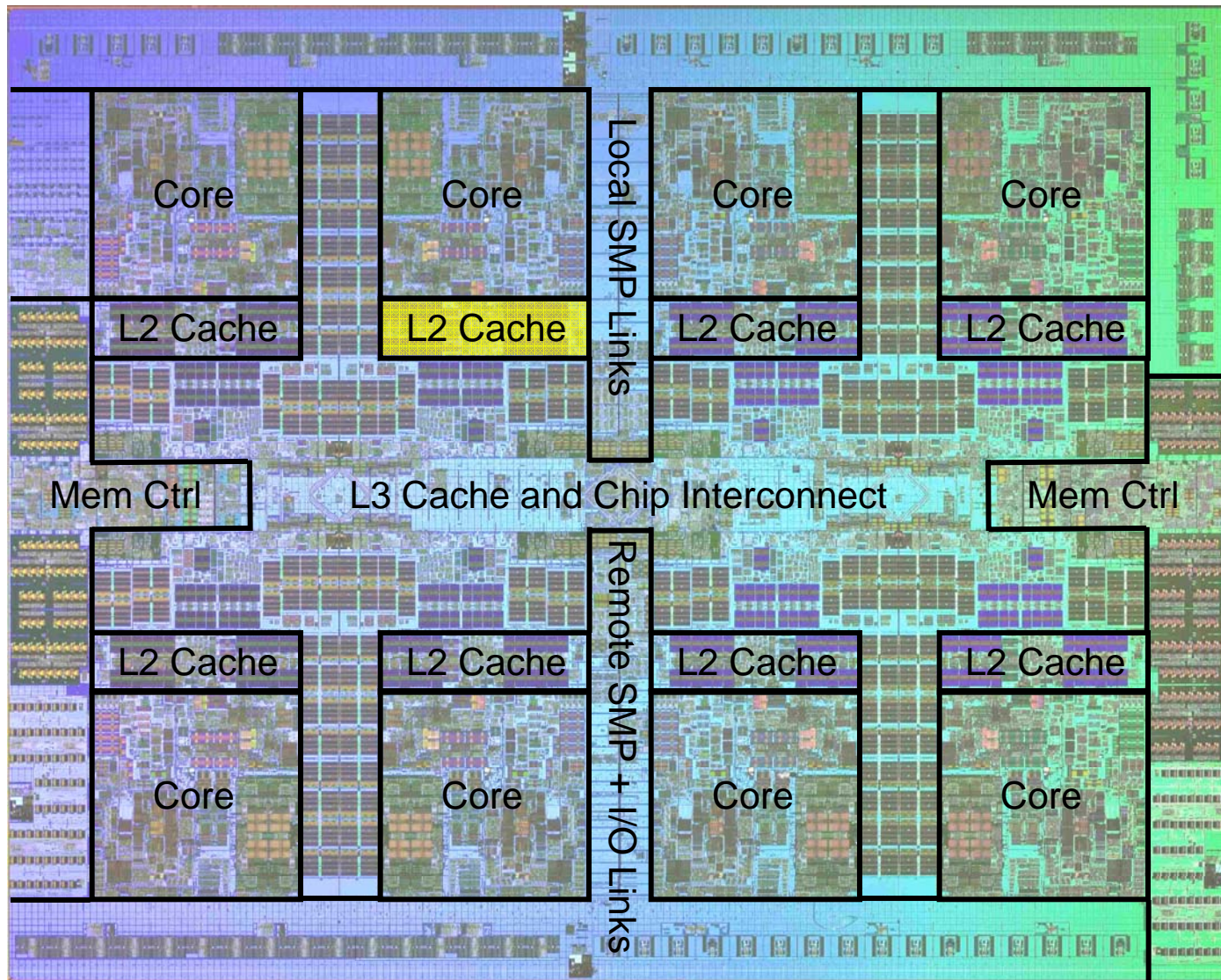
### Solution: L2 “Turbo” Cache



- L2 “Turbo” cache keeps a tight 256K working set with extremely low latency (~3X lower than local L3 region) and high bandwidth, reducing L3 power and boosting performance.

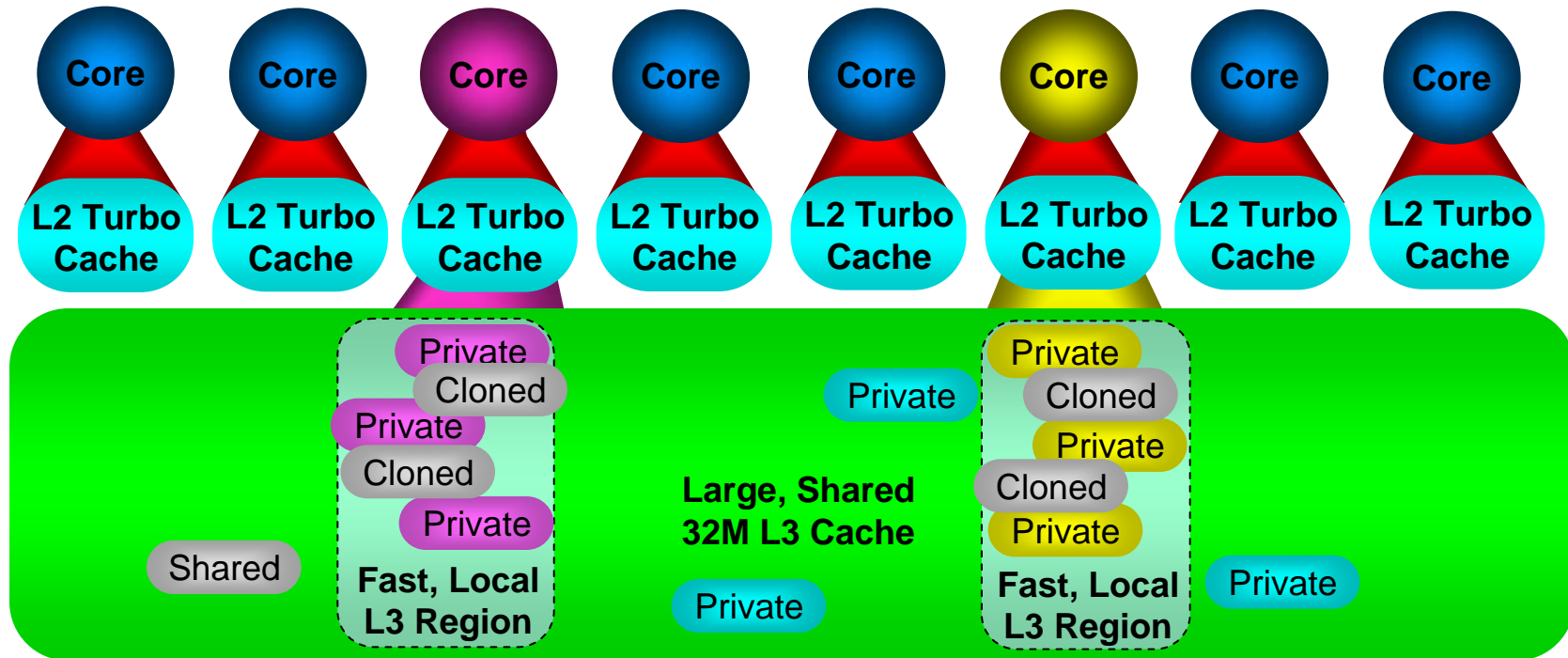


# Cache Hierarchy Technology and Innovation



# Cache Hierarchy Technology and Innovation

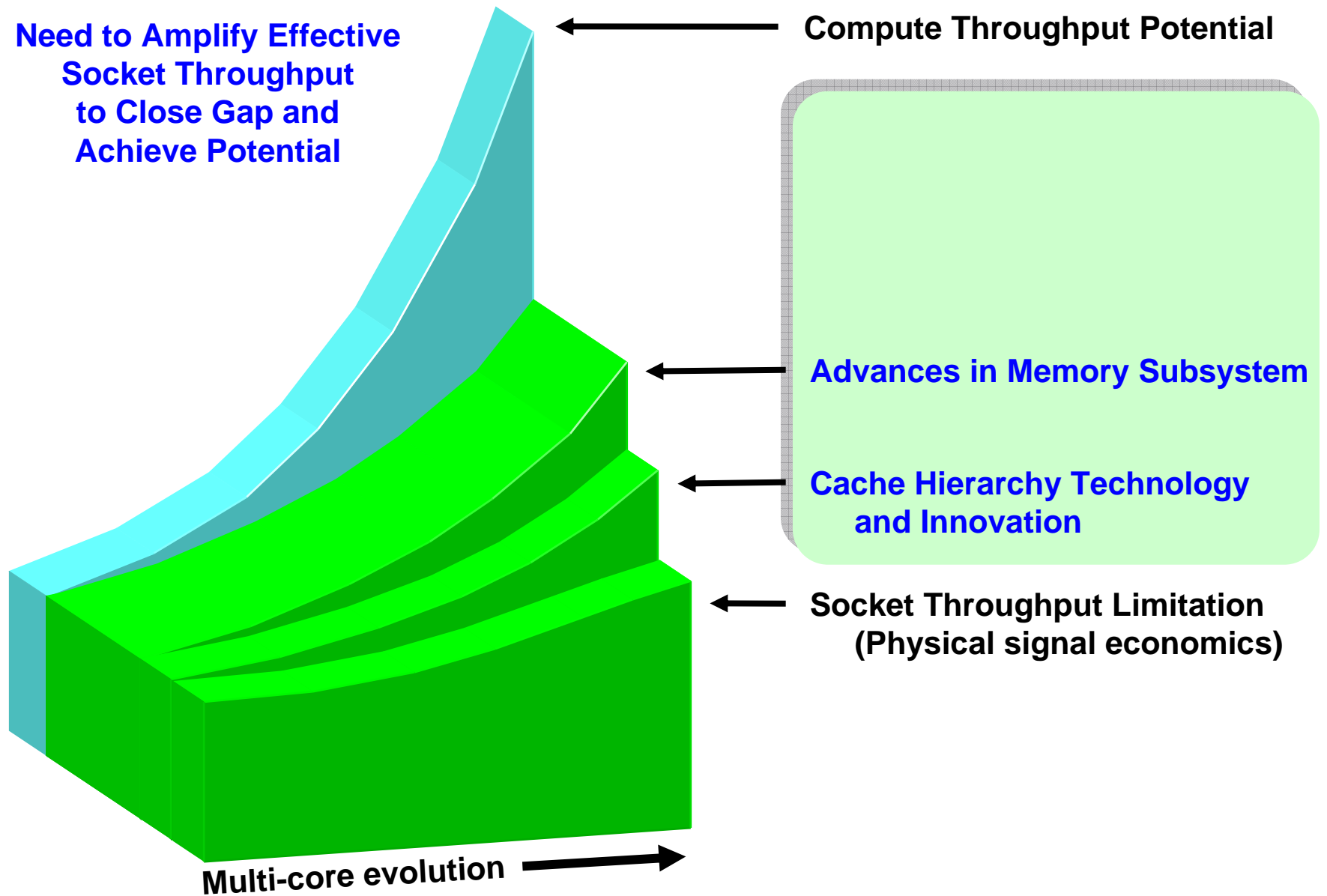
## Cache Hierarchy Summary



Cache Level	Capacity	Array	Policy	Comment
L1 Data	32K	Fast SRAM	Store-thru	Local thread storage update
Private L2	256K	Fast SRAM	Store-In	De-coupled global storage update
Fast L3 Region	Up to 4M	eDRAM	Partial Victim	Reduced power footprint (up to 4M)
Shared L3	32M	eDRAM	Adaptive	Large 32M shared footprint

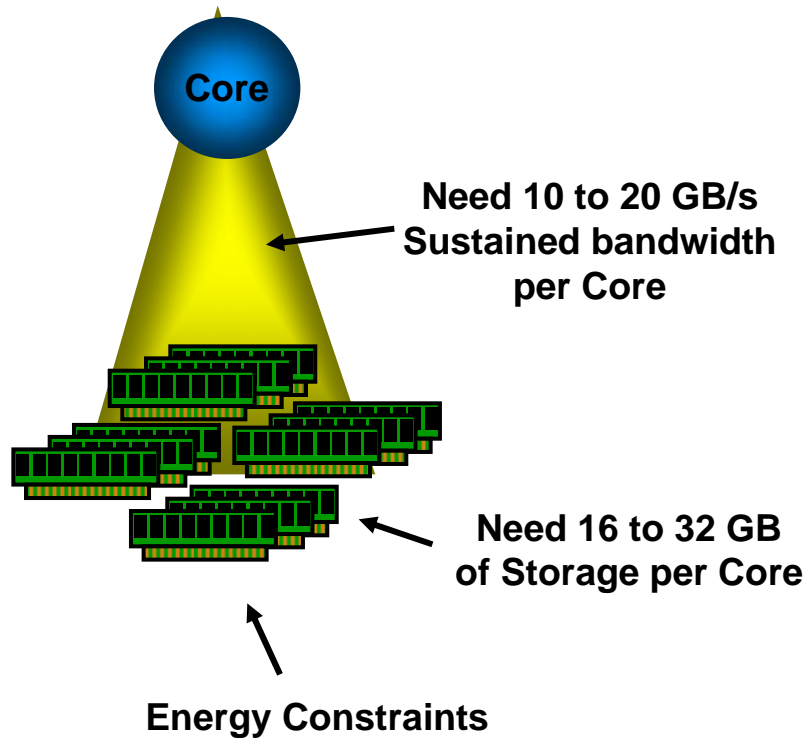
# Challenge: How does POWER7 maintain the Balance?

**Need to Amplify Effective Socket Throughput to Close Gap and Achieve Potential**



## Advances in Memory Subsystem

### Memory Subsystem Rqmt for POWER Servers



### Challenge for Multi-core POWER7

#### Socket Challenge:

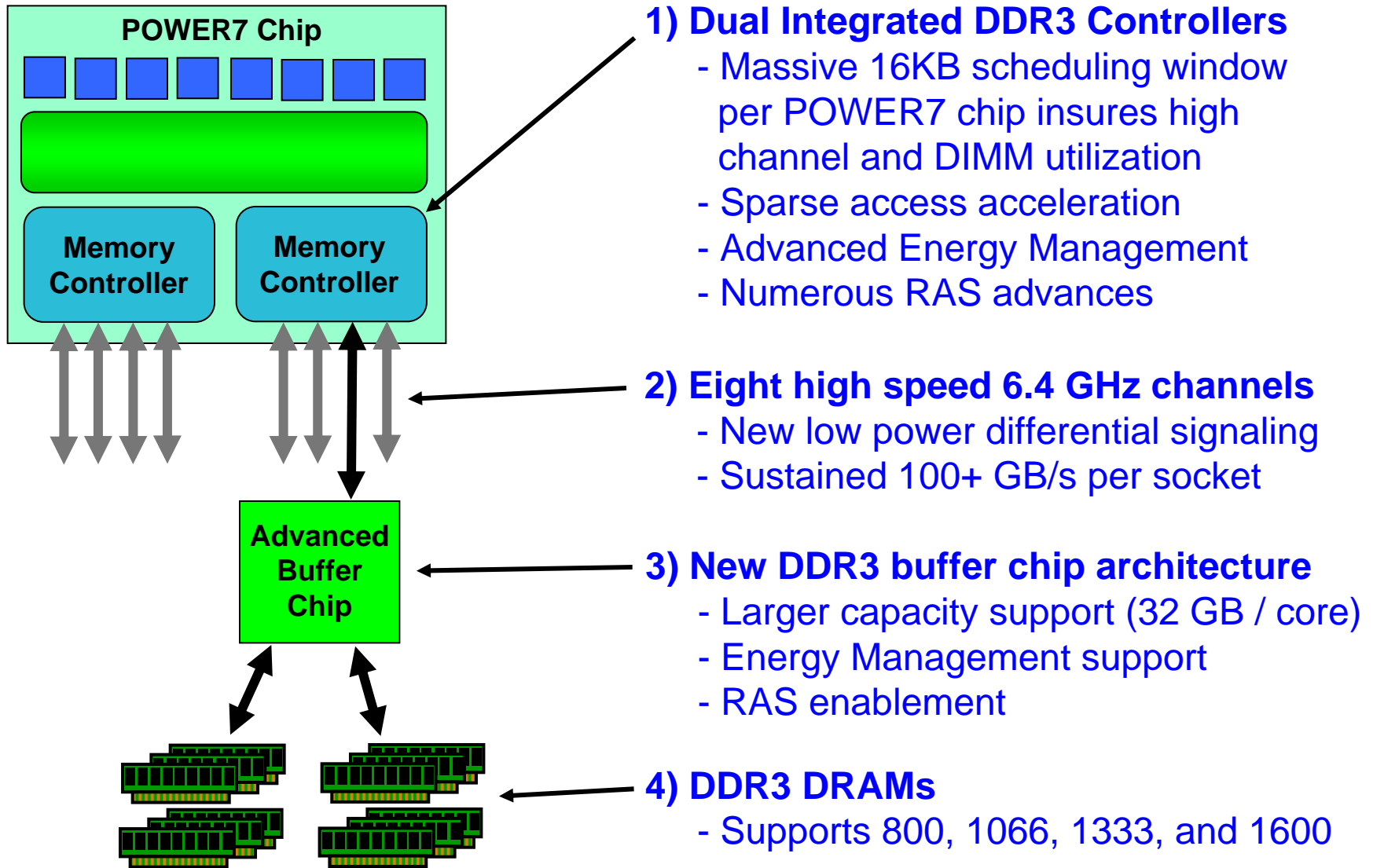
4x growth in memory bandwidth and capacity needed per socket.

#### System Challenge:

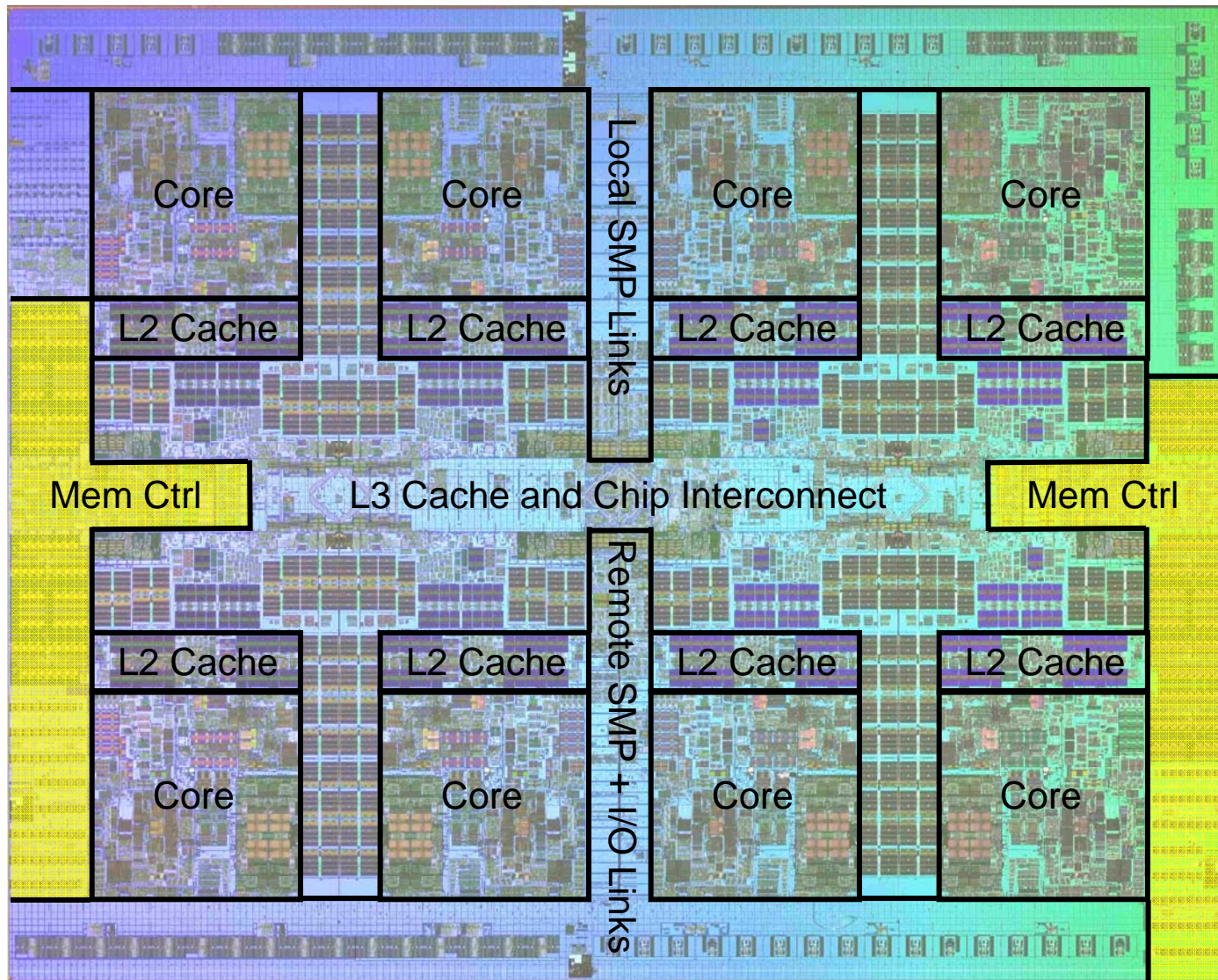
Packaging more memory into similar volume with similar energy and cooling constraints.

## Advances in Memory Subsystem

### Multi-faceted Solution

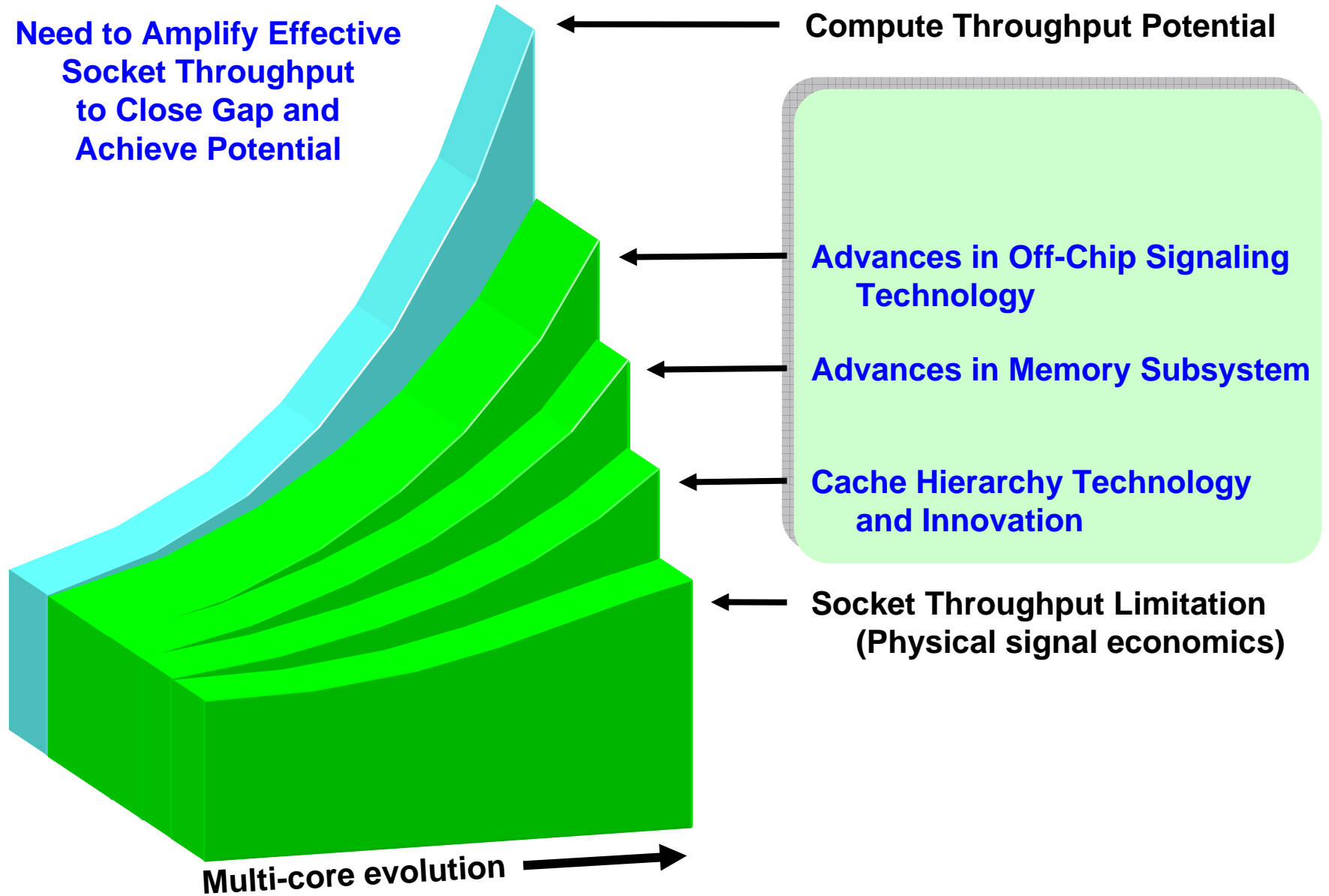


## Advances in Memory Subsystem



# Challenge: How does POWER7 maintain the Balance?

**Need to Amplify Effective Socket Throughput to Close Gap and Achieve Potential**



## Advances in Off-chip Signaling Technology

- 1) Enhanced Signal-ended “Elastic Interface” Technology
- 2) New high speed, low power Differential Technology

Interface	Signal Type	Info Width	Frequency	Bandwidth
Off-chip Cache	none	none	none	none
Memory Channels	Differential	28 bytes	6.4 Ghz	180 GB/s
I/O Bridge	Single-ended	20 bytes	2.5 Ghz	50 GB/s
SMP Interconnect	Single-ended	120 bytes	3.0 Ghz	360 GB/s
Total Bandwidth				590 GB/s

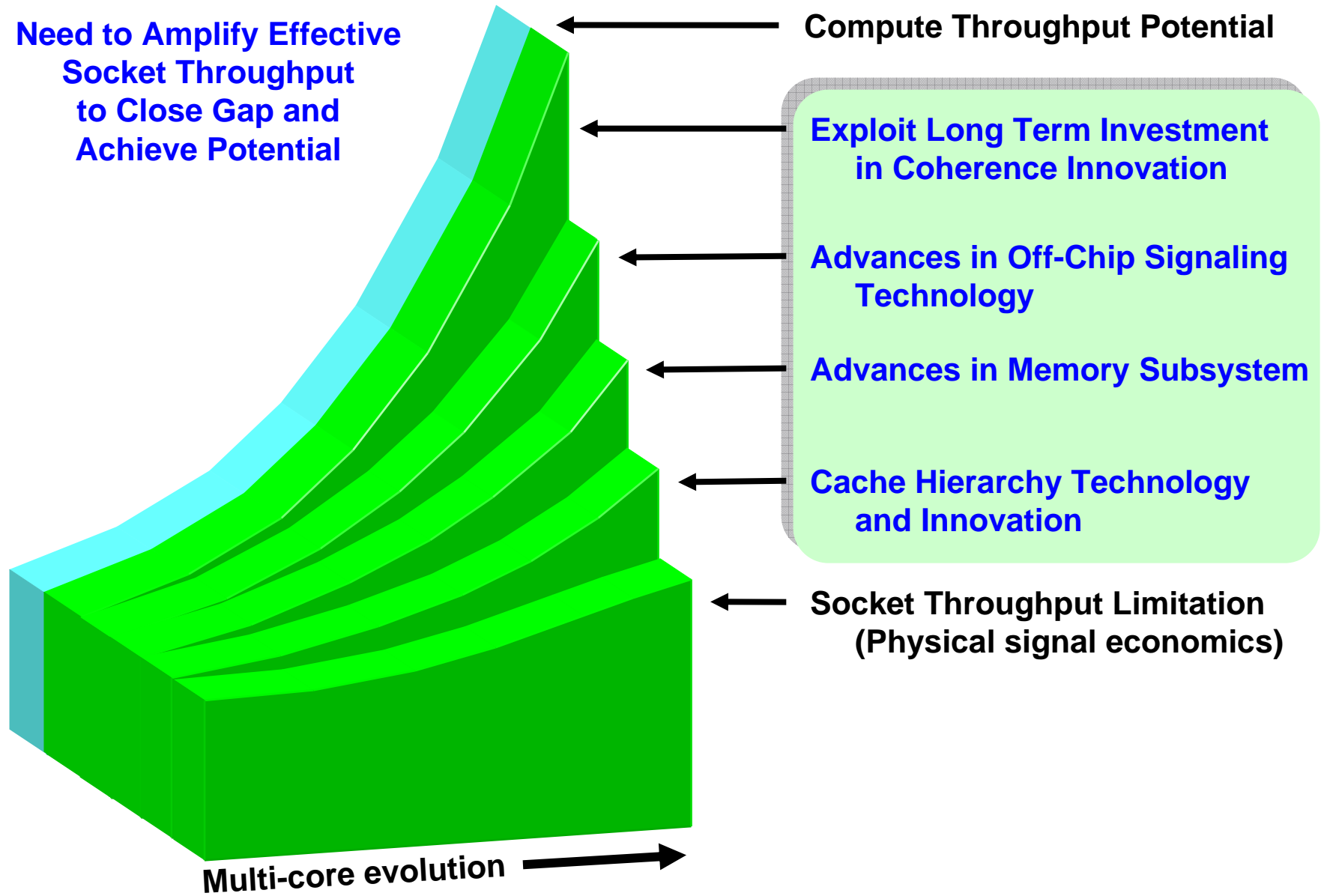
(Note that bandwidths shown are raw, peak signal bandwidths)

- Moving L3 onto POWER7 along with advances in signaling technology enables significant raw bandwidth growth for both memory and I/O subsystems. Note that advanced scheduling improves POWER7’s ability to utilize memory bandwidth.

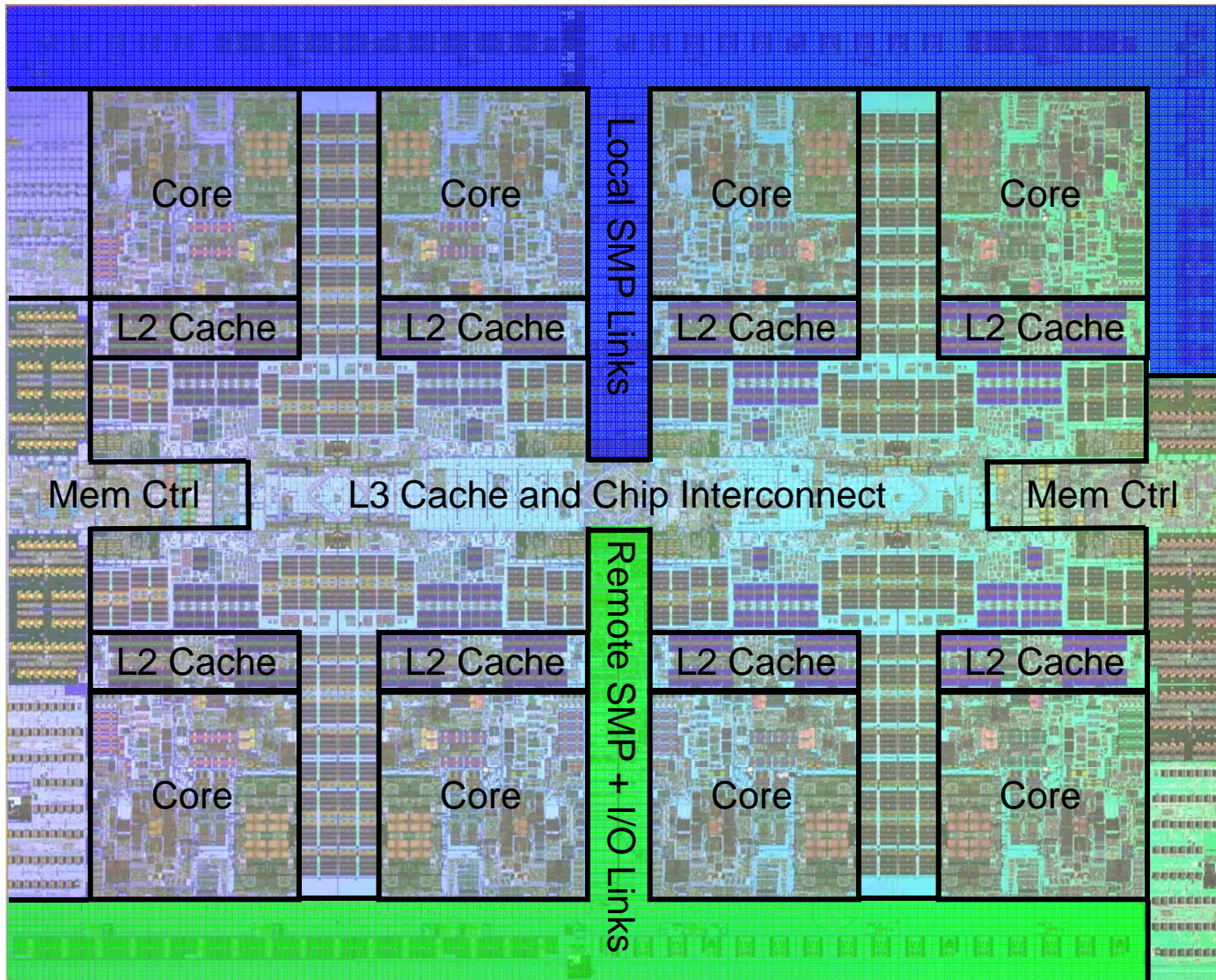


# Challenge: How does POWER7 maintain the Balance?

**Need to Amplify Effective Socket Throughput to Close Gap and Achieve Potential**

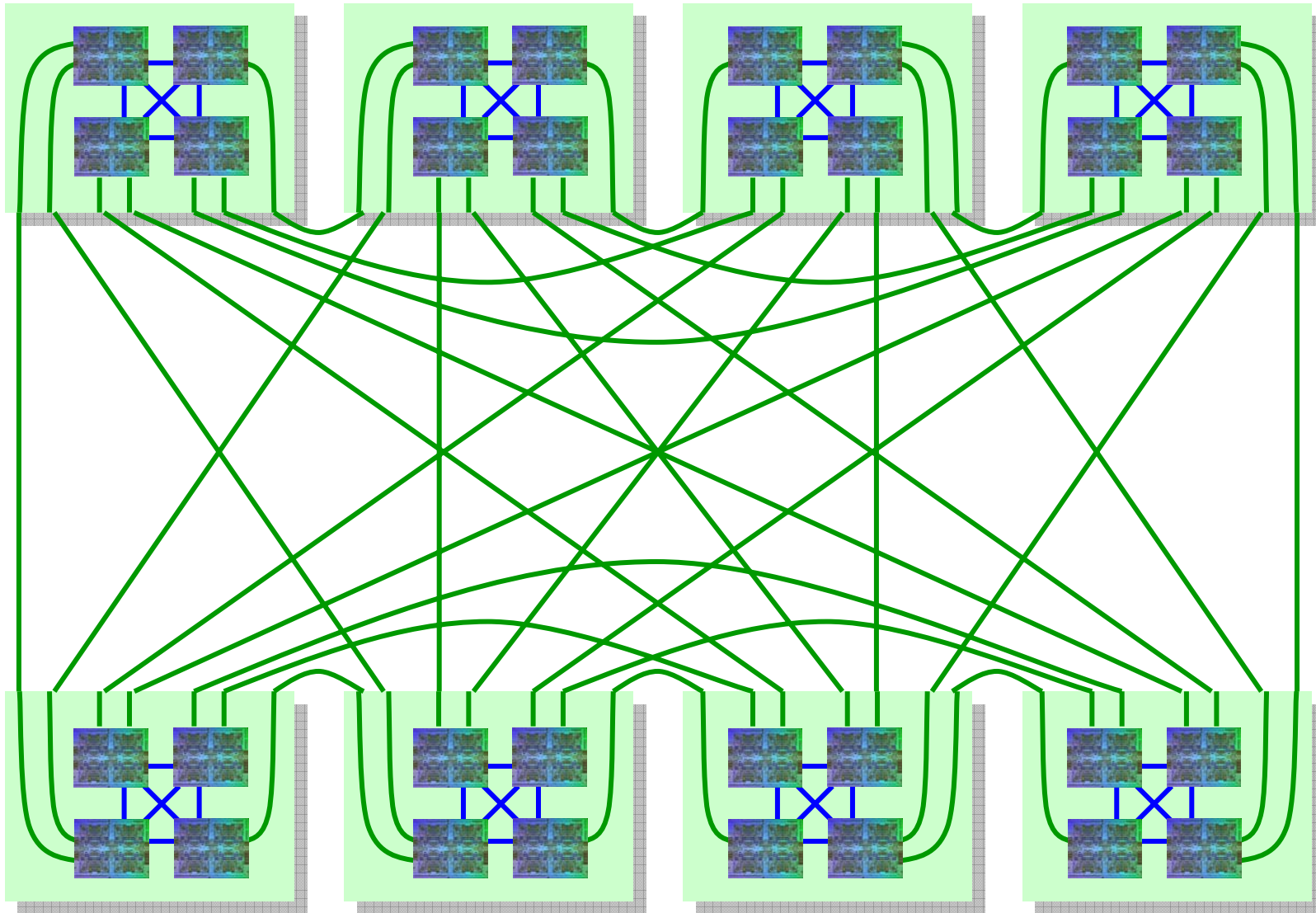


## Exploit Long Term Investment in Coherence Innovation



Using local and remote SMP links, up to 32 POWER7 chips are connected

## Exploit Long Term Investment in Coherence Innovation



**Up to 32 POWER7 chips form a massive SMP system.**

\* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.

## Exploit Long Term Investment in Coherence Innovation

### Coherence Protocol Features

- POWER storage Architecture enables decoupled global storage updates. Updates can be reordered and are effectively “deserialized”.
- Decentralized coherence resolution, and bounded latency broadcast transport layer.
- Decentralized coherence resolution, advanced cache states, optimized on-chip transport, and broadcast free barriers.

### POWER7 Exploitation

- POWER Servers can drive massive coherence throughput. A 32-chip POWER7 system can manage over 20,000 concurrently reordered coherent storage operations (~4X more than POWER6 systems), with minimal tracking overhead per operation.
- Low latency intervention, high performance locking constructs, and robust scaling.

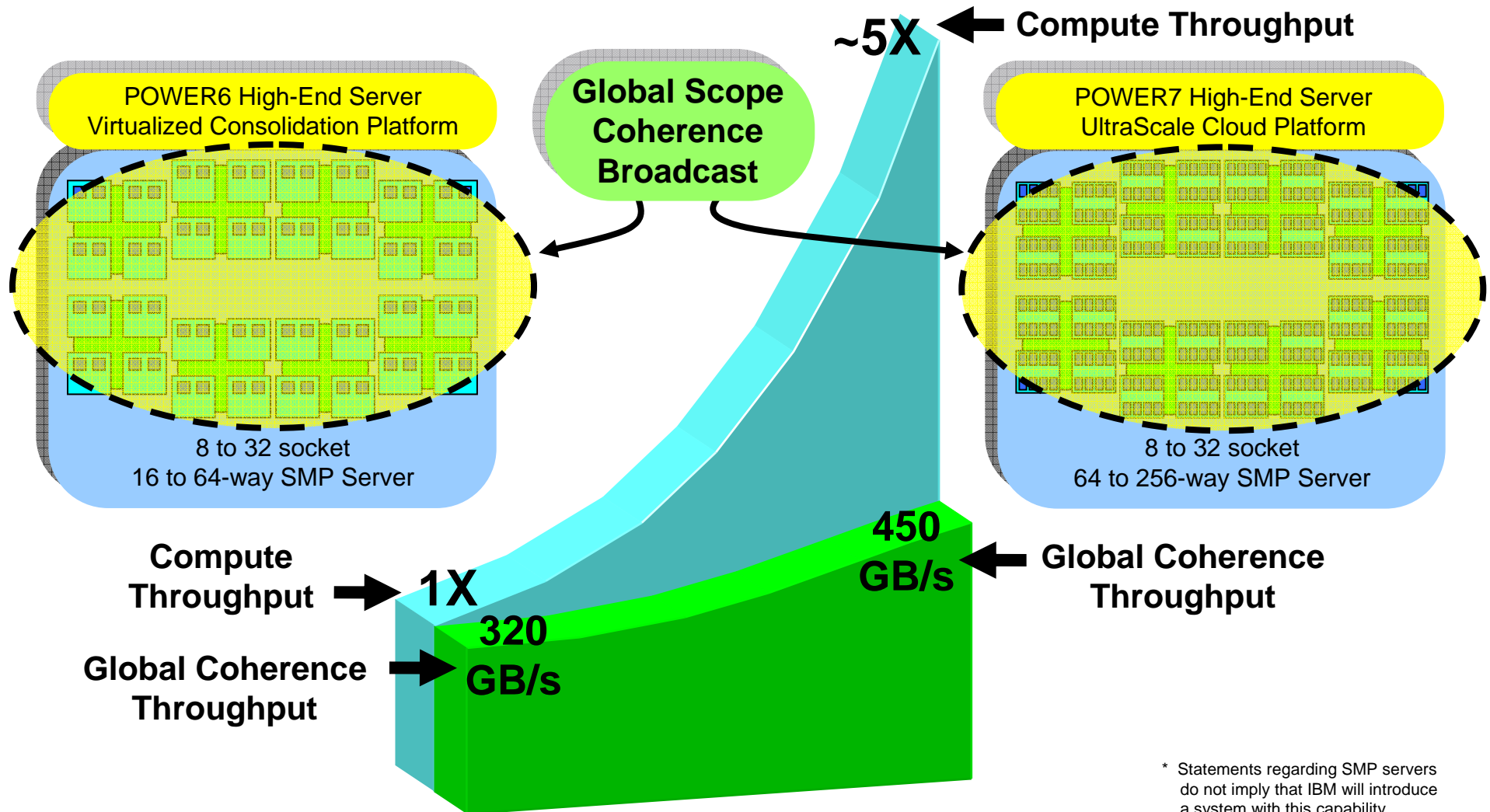
### Key Ingredients for Balanced Scaling in Traditional POWER Servers:

- Architecture enables re-ordered, decoupled storage updates
- Decentralized coherence resolution
- Broadcast transport layer

\* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.

# Exploit Long Term Investment in Coherence Innovation

Challenge: As system size grows, Coherence broadcast traffic increases

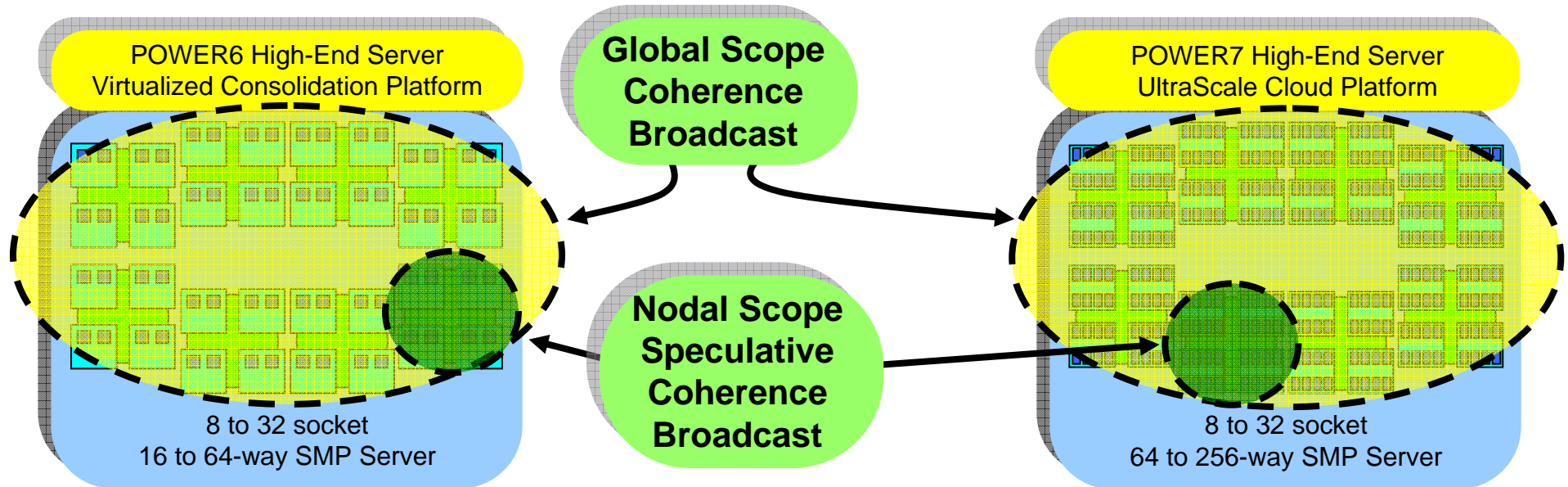


\* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.

## Exploit Long Term Investment in Coherence Innovation

### Solution: Speculative limited scope Coherence broadcast

- In 2003, recognized emerging trend
- Developed Dual-Scope Broadcast Coherence Protocol for POWER6
- Utilizes 13 cache states and integrated scope indicator in memory



### Provides value for POWER6

- Latency reduction
- Near Perfect Scaling for extreme memory intensive workloads
- Ultra-dense packaging (Power 575)

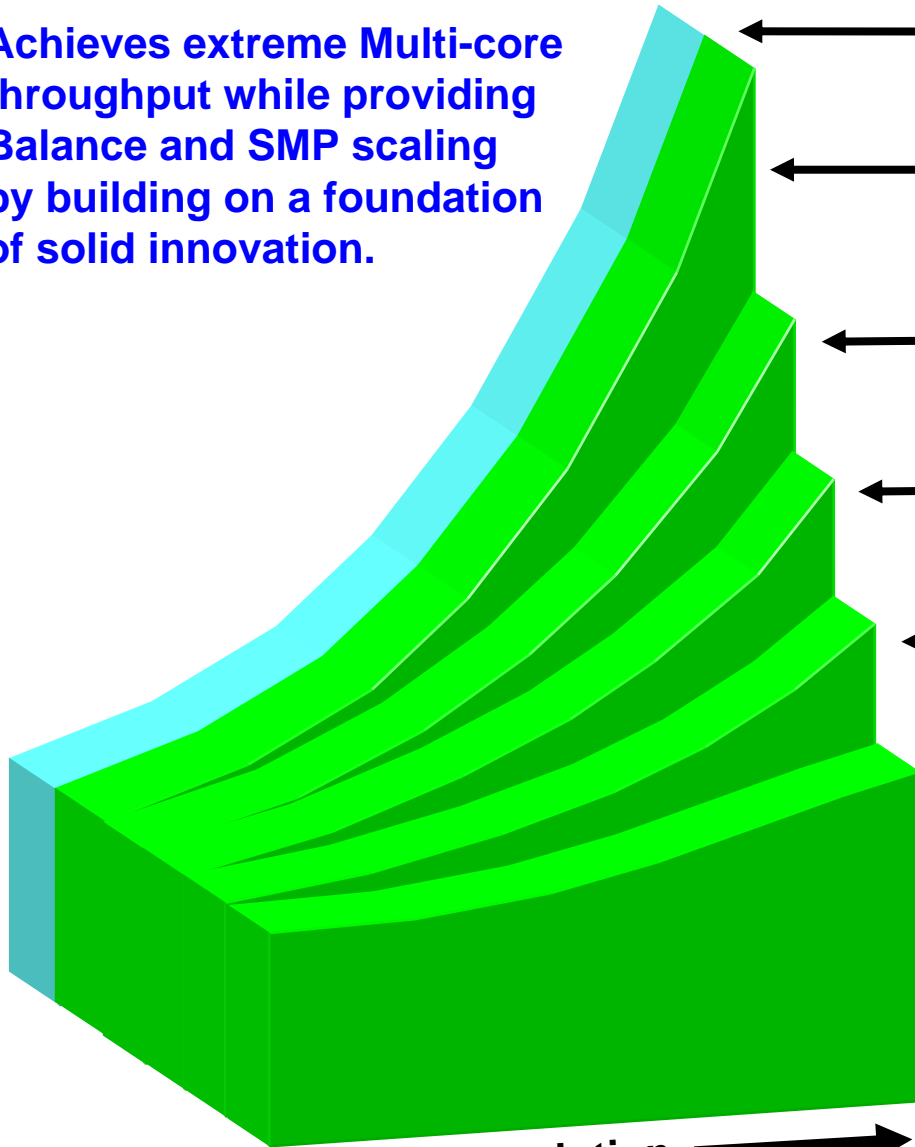
### Necessity for POWER7

- 450 GB/s must grow to 1.6 TB/s to match POWER6 scaling
- 450 GB/s → 3.6 TB/s theoretical peak
- 3.6 TB/s → 14.4 TB/s with chip scope

\* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.

## Summary: POWER7 maintains the Balance

Achieves extreme Multi-core throughput while providing Balance and SMP scaling by building on a foundation of solid innovation.



← Compute Throughput Potential

← Exploit Long Term Investment in Coherence Innovation

← Advances in Off-Chip Signaling Technology

← Advances in Memory Subsystem

← Cache Hierarchy Technology and Innovation

← Socket Throughput Limitation (Physical signal economics)

IBM POWER chips uniquely positioned to excel given the emerging trends:

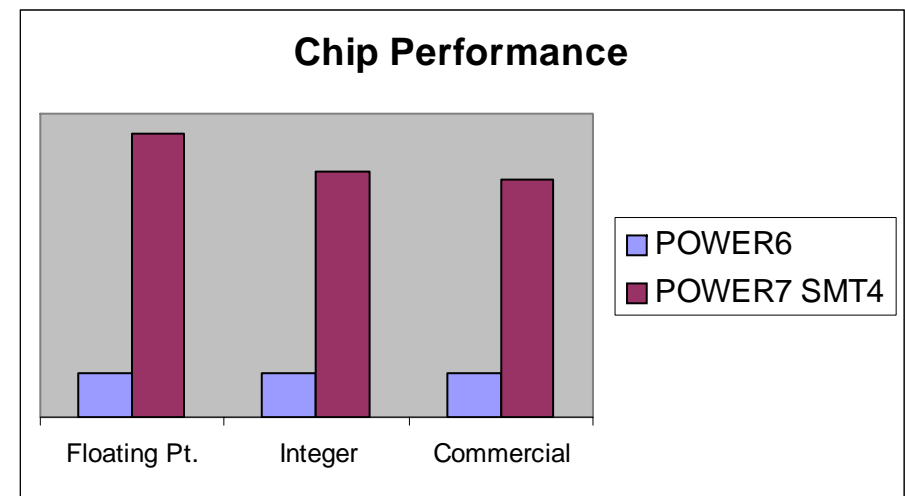
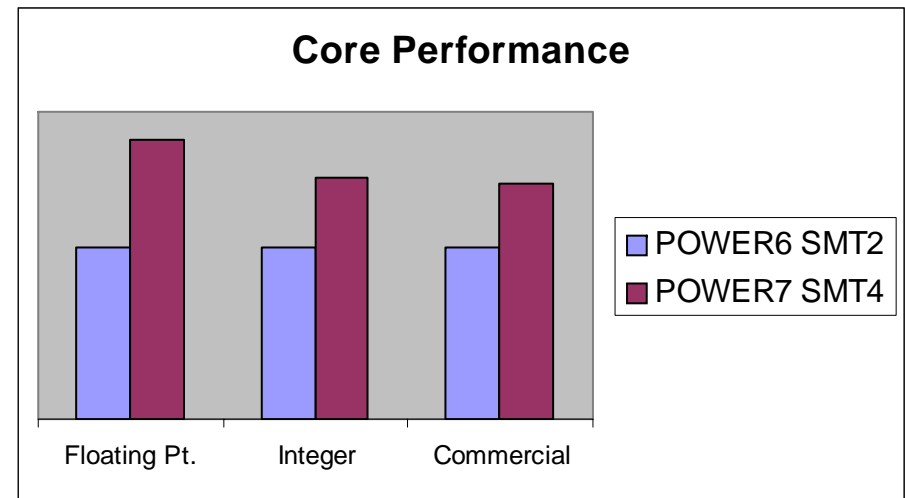
- 1) History of large SMP leadership
- 2) Storage Architecture economics
- 3) High density packaging leadership

Multi-core evolution →

## POWER7: Performance Estimates

### POWER7 Continues Tradition of Excellent Scalability

- Core performance increased by:
  - Re-pipelined execution units
  - Reduced L1 cache latency
  - Tightly coupled L2 cache
  - Additional execution units
  - More flexible execution units
  - Increased pipeline utilization with SMT4 and aggressive out of order execution
  
- Chip Performance Improved Greater than 4X:
  - High performance on chip interconnect
  - Improved storage architecture
  - Dual high speed integrated memory controllers
  
- System
  - Achieves extreme Multi-core throughput while providing Balance and SMP scaling by building on a foundation of solid innovation
  - Advanced SMP links will provide near linear scaling for larger POWER7 systems.



\* Performance estimates relate to processor only and should not be used to estimate projected server performance.

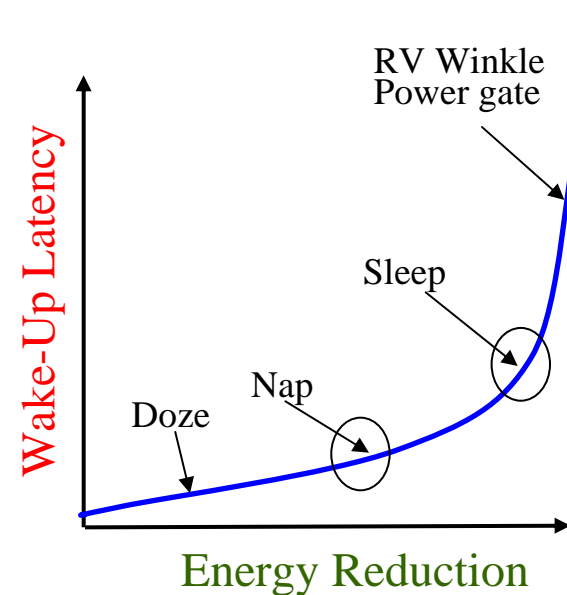


## Energy Management: Architected Idle Modes

### Two Design Points Chosen for Technology

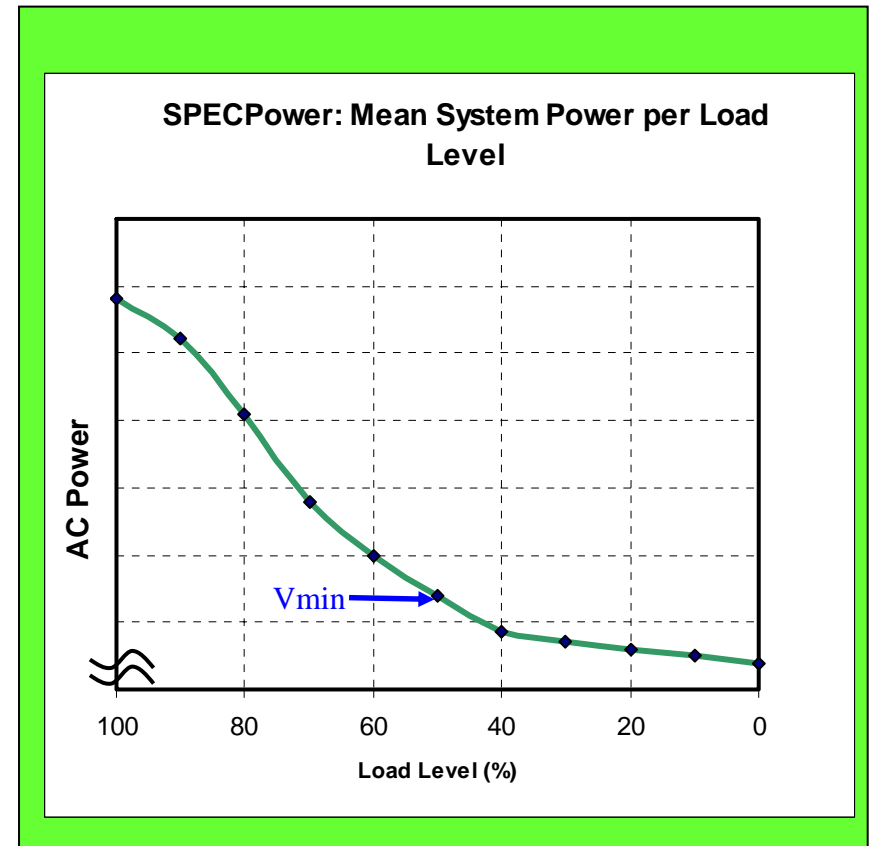
- Nap (optimized for wake-up time)
  - Turn off clocks to execution units
  - Reduce frequency to core
  - Caches and TLB remain coherent
  - Fast wake-Up
- Sleep (optimized for power reduction)
  - Purge caches and TLB
  - Turn off clocks to full core and caches
  - Reduce voltage to V-retention
    - Leakage current reduced substantially
  - Voltage ramps-up on wake up
  - No core re-initialization required

### 4 PowerPC Architected States



## Adaptive Energy Management: Energy Scale™

- Chip FO4 Tuned for Optimal Performance/Watt in Technology
- DVFS (Dynamic Voltage and Frequency Slewing)
  - -50% to +10% frequency slew independent per core
  - Frequency and voltage adjusted based on:
    - Work load and utilization.
    - On board activity monitors
- Turbo-Mode
  - Up to 10% frequency boost
  - Leverages excess energy capacity from:
    - Non worst case work loads
    - Idle cores
- Processor and Memory Energy Usage can be independently Balanced.
  - Real time hardware performance monitors used.
  - On board power proxy logic estimates power
- Power Capping Support
  - Allows budgeting of power to different parts of system

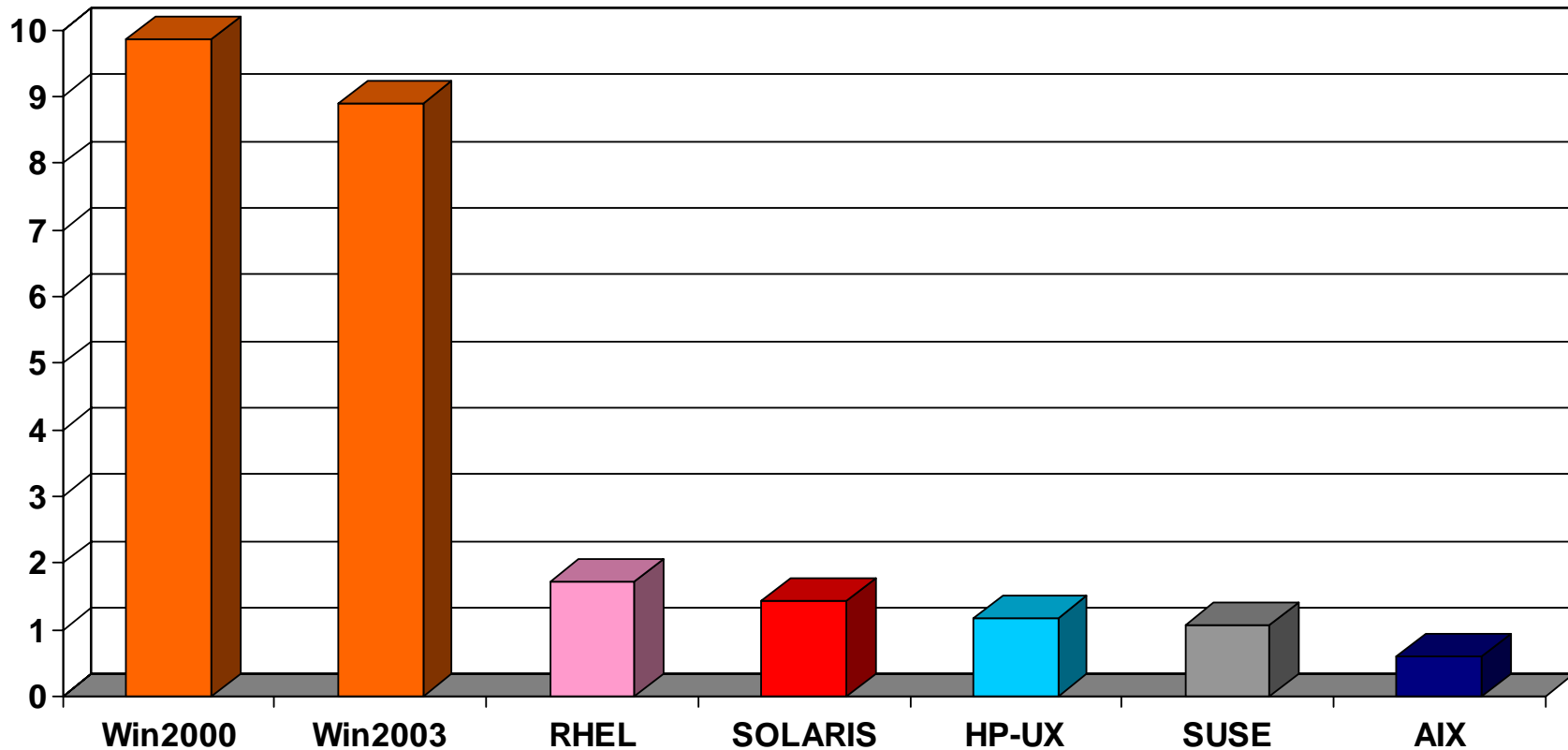


## Power Systems – Reliability, Availability, Serviceability (RAS)

# OS Downtime Comparison Survey

**400 participants in 27 countries**

Hours



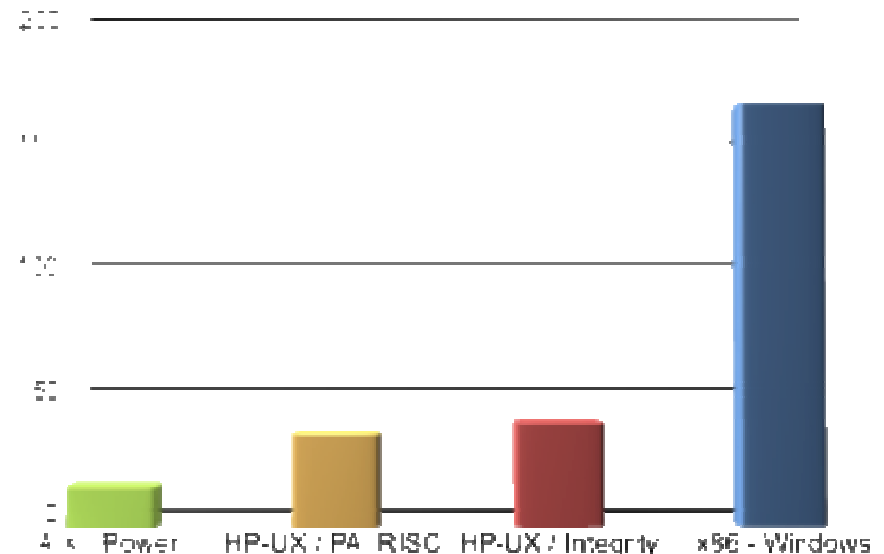
The Yankee Group “2007-2008 Global Server Operating Systems Reliability Survey” as quoted in “Windows Server: The New King of Downtime” by Mark Joseph Edwards at [www.windowssitpro.com/article/articleid/98475/windows-server-the-new-king-of-downtime.html](http://www.windowssitpro.com/article/articleid/98475/windows-server-the-new-king-of-downtime.html), March 5, 2008 and in <http://www.sunbeltsoftware.com/stu/Yankee-Group-2007-2008-Server-Reliability.pdf>

## ITIC Survey says Power Systems with AIX deliver 99.997% uptime

- 54% of IT executives and managers say that they require 99.99% or better availability for their applications

- Power Systems with AIX delivers the best RAS of UNIX, Linux, Windows choices
  1. **Availability: The least amount of downtime**
    - 15 minutes a year
    - 2.3 times better than the closest UNIX competitor
    - more than 10X better than Windows
  2. **Reliability: The fewest unscheduled outages**
    - less than one outage per year
  3. **Serviceability: The fastest patch time**
    - 11 minutes to apply a patch

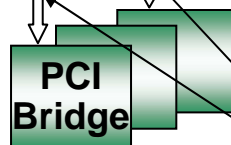
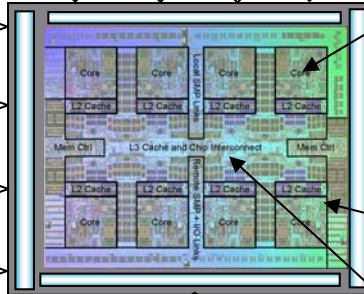
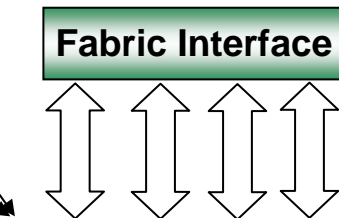
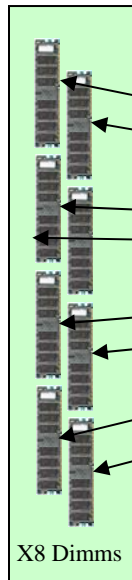
Minutes of downtime per year



Source: Network World, dated July 14, 2009, reports on the 2009 ITIC Global Server Hardware & Server OS Reliability Survey Results

# POWER7: Reliability and Availability Features

## Dynamic Oscillator Failover



## Fabric Bus Interface to other Chips and Nodes

- ECC protected
- Node hot add /repair

## Core Recovery

- Leverage speculative execution resources to enable recovery
- Error detected in GPRs FPRs VSR, flushed and retried
- Stacked latches to improve SER

## Alternate Processor Recovery

- Partition isolation for core checkstops

## L3 eDRAM

- ECC protected
- SUE handling
- Line delete
- Spare rows and columns

## GX IO Bus

- ECC protected
- Hot add

## InfiniBand® Interface

- Redundant paths

- 64 Byte ECC on Memory
  - Corrects full chip kill on X8 dimms
  - Spare X8 devices implemented
- Dual memory chip failures do not cause outage
- Selective memory mirror capability to recover partition from dimm failures
- Hardware assisted scrubbing
- SUE handling
- Dynamic sparing on channel interface
- PowerVM Hypervisor protected from full DIMM failures

\* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.

## Power Systems Benefits

- IBM Power Systems have a consistent, reliable history of executing on schedule allowing customers to confidently plan for the future
- IBM Power Systems offer highest performance reducing the need for additional resources
- IBM Power Systems are designed for performance with high reliability and availability
  - Moving towards Continuous Availability – hardware and software
  - Reduced and shorter outages lower costs and improve SLAs
- Virtualization capabilities intrinsic to Power Systems design allows improved service and lower costs by consolidating
  - POWER7 systems increased to up to 1000 partitions / system
  - POWER7 systems designed to leverage, exploit and enhance current PowerVM capabilities

## Summary

Power Systems™ continue strong

- 7th Generation Power chip:
  - Balanced Multi-Core design
  - EDRAM technology
  - SMT4
- Greater than 4X performance in same power envelope as previous generation
- Scales to 32 socket, 1024 threads balanced system
- Building block for peta-scale PERCS project
- Achieves extreme Multi-core throughput while providing Balance and SMP scaling by building on a foundation of solid innovation



*Power7 High Volume Card*

POWER7 Systems Running in Lab

- AIX®, IBM i, Linux® all operational

\* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.





# POWER7 Processors: The Beat Goes On

Joel M. Tandler, Executive IT Architect  
[jtandler@us.ibm.com](mailto:jtandler@us.ibm.com)

Acknowledgment: This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0002



## Trademarks

**The following are trademarks of the International Business Machines Corporation in the United States, other countries, or both.**

Not all common law marks used by IBM are listed on this page. Failure of a mark to appear does not mean that IBM does not use the mark nor does it mean that the product is not actively marketed or is not significant within its relevant market.

Those trademarks followed by ® are registered trademarks of IBM in the United States; all others are trademarks or common law marks of IBM in the United States.

For a complete list of IBM Trademarks, see [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml):

\*, AS/400®, e business (logo)®, DBE, ESCO, eServer, FICON, IBM®, IBM (logo)®, iSeries®, MVS, OS/390®, pSeries®, RS/6000®, S/30, VM/ESA®, VSE/ESA, WebSphere®, xSeries®, z/OS®, zSeries®, z/VM®, System i, System i5, System p, System p5, System x, System z, System z9®, BladeCenter®

**The following are trademarks or registered trademarks of other companies.**

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

\* All other products may be trademarks or registered trademarks of their respective companies.

### Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.