

# Roadrunner System Overview

October 17, 2007

(slides from Roadrunner Technical Assessment)

Ken Koch

Roadrunner Technical Manager,  
Computer, Computational, and Statistical Sciences Division,  
Los Alamos National Laboratory

Work presented was performed by a large team of Roadrunner project staff!

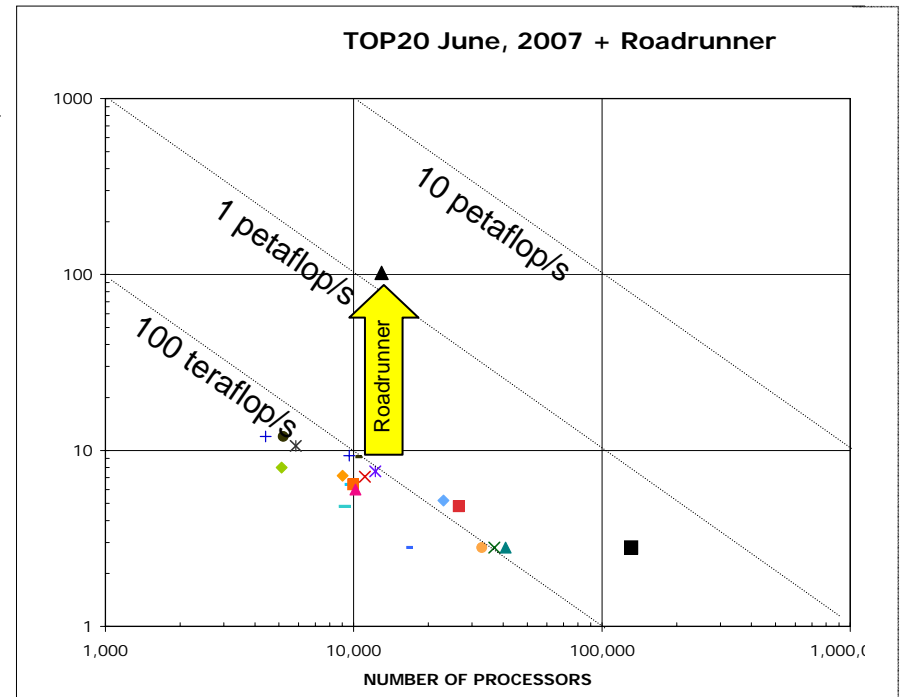
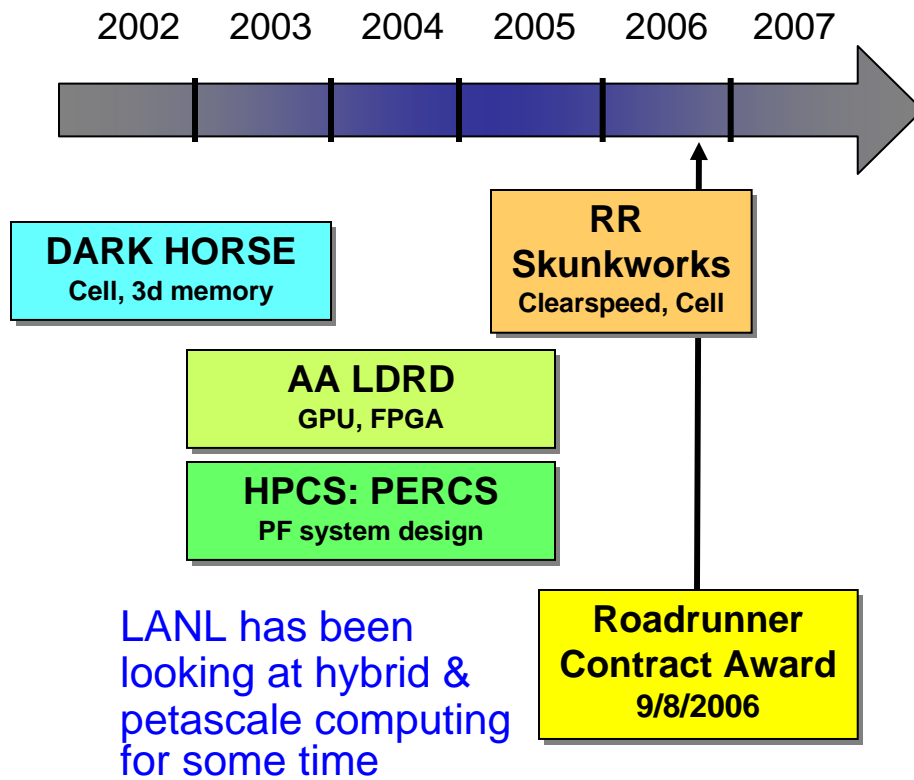


EST. 1943  
Operated by the Los Alamos National Security, LLC for the DOE/NNSA

LA-UR-07-7235

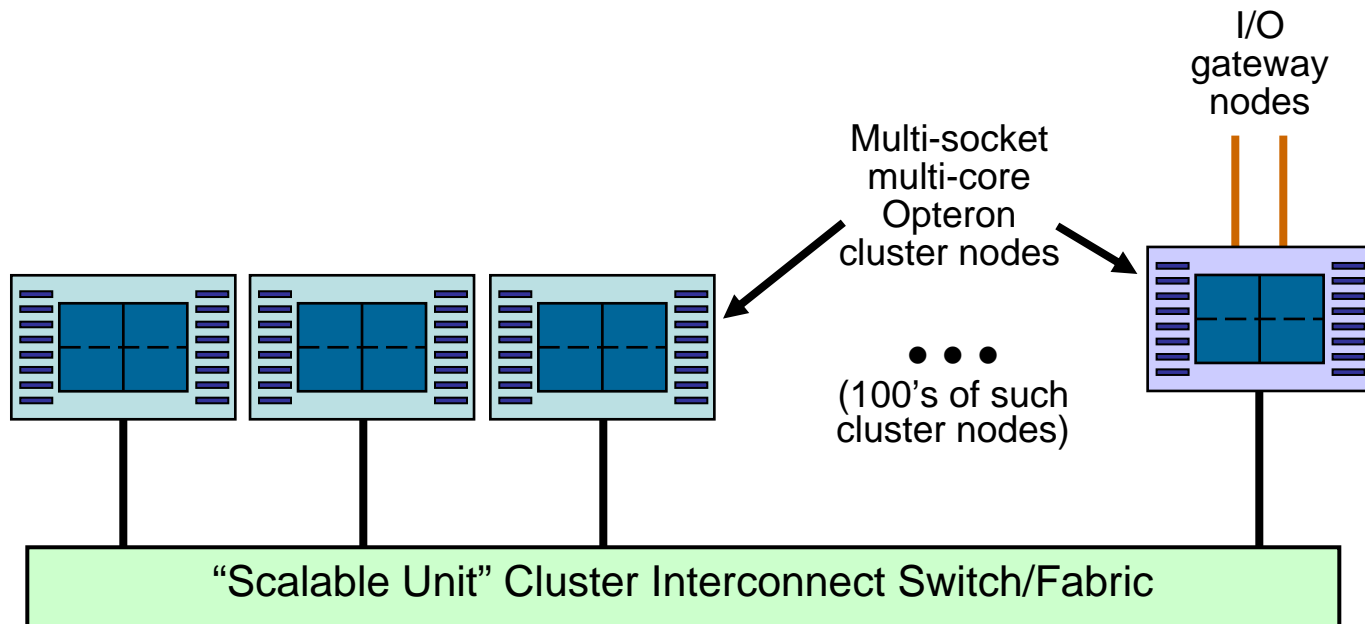


# Hybrid computing is a transformational technology

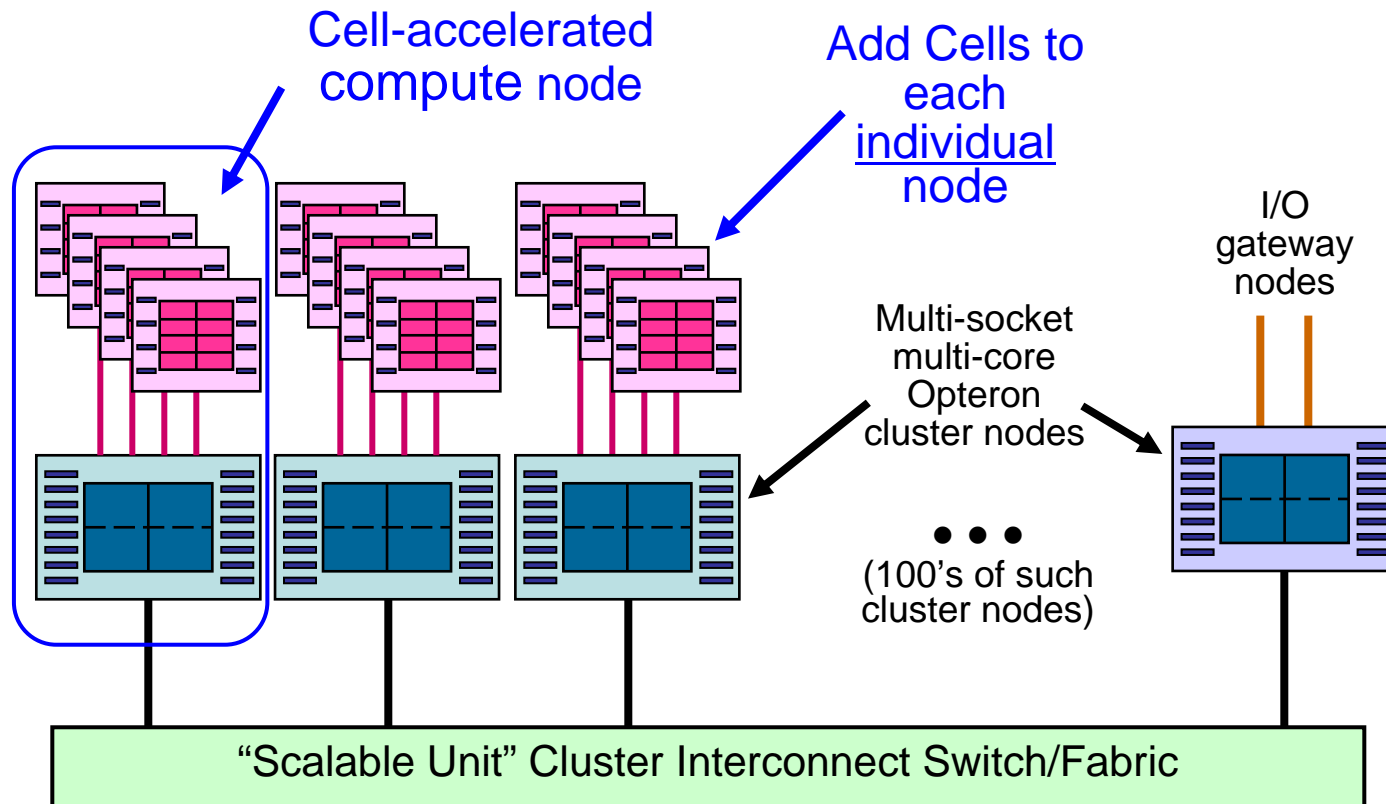


Roadrunner is on a different path to a petascale system

# Roadrunner Phase 3 is Cell-accelerated, not a cluster of Cells



# Roadrunner Phase 3 is Cell-accelerated, not a cluster of Cells

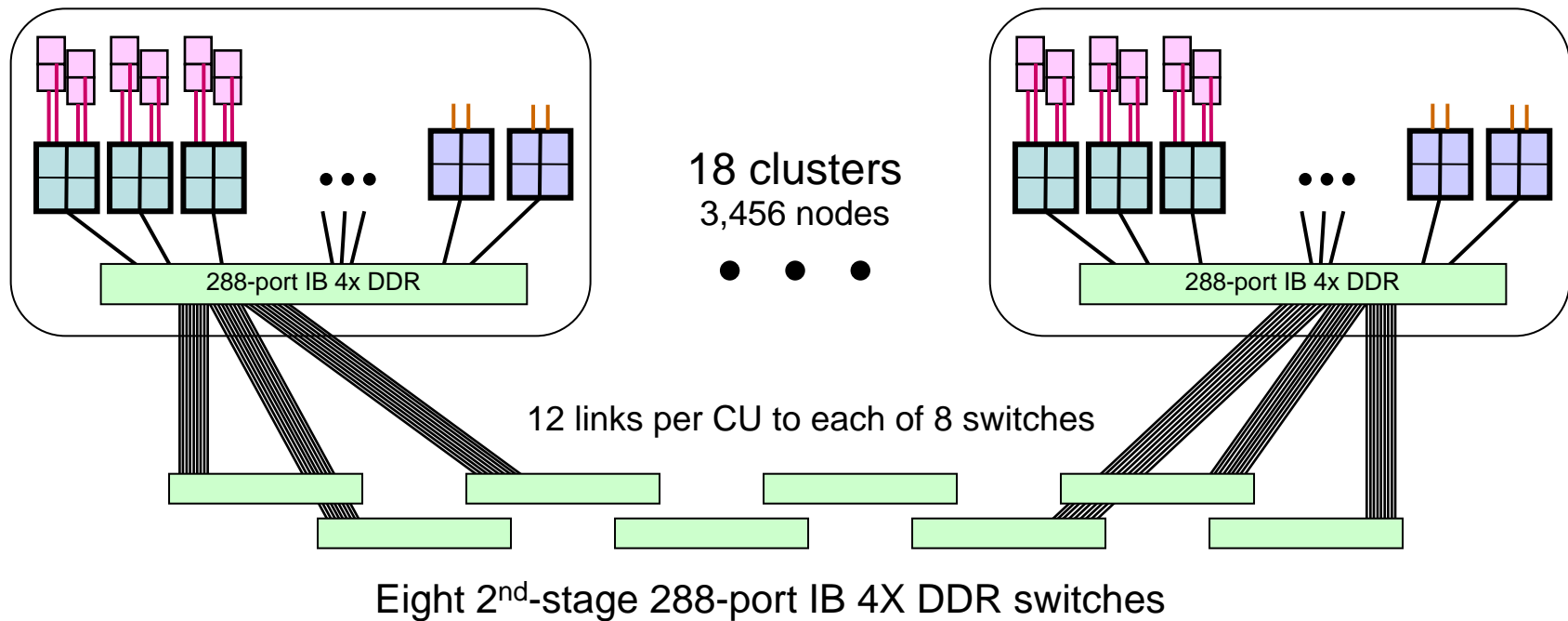


Node-attached Cells is what makes Roadrunner different!

# Roadrunner is a hybrid petascale system of modest size delivered in 2008

**Connected Unit cluster**  
180 compute nodes w/ Cells  
12 I/O nodes

6,912 dual-core Optrons  $\Rightarrow$  50 TF  
12,960 Cell eDP chips  $\Rightarrow$  1.3 PF



# Roadrunner Phase 3 has changed for the better since last winter

---

- IBM & LANL redesigned the Cell-accelerated compute node and clusters to improve performance
  - *Cell-Opteron links:*
    - one IB SDR  $\Rightarrow$  two PCIe x8
    - 4x the BW & better latency
  - *Cluster Nodes:*
    - IB SDR  $\Rightarrow$  IB DDR with half-sized nodes
    - 4x the node BW & better latency
  - *Hybrid node:*
    - 4U server + blades  $\Rightarrow$  integrated blade packaging (Triblade)
    - Same compute density & heat density per rack
  - *Connected clusters*
    - Full-BW at 1st stage and half-BW at 2nd stage
    - Full-system bi-section BW per Cell chip is ~2x better

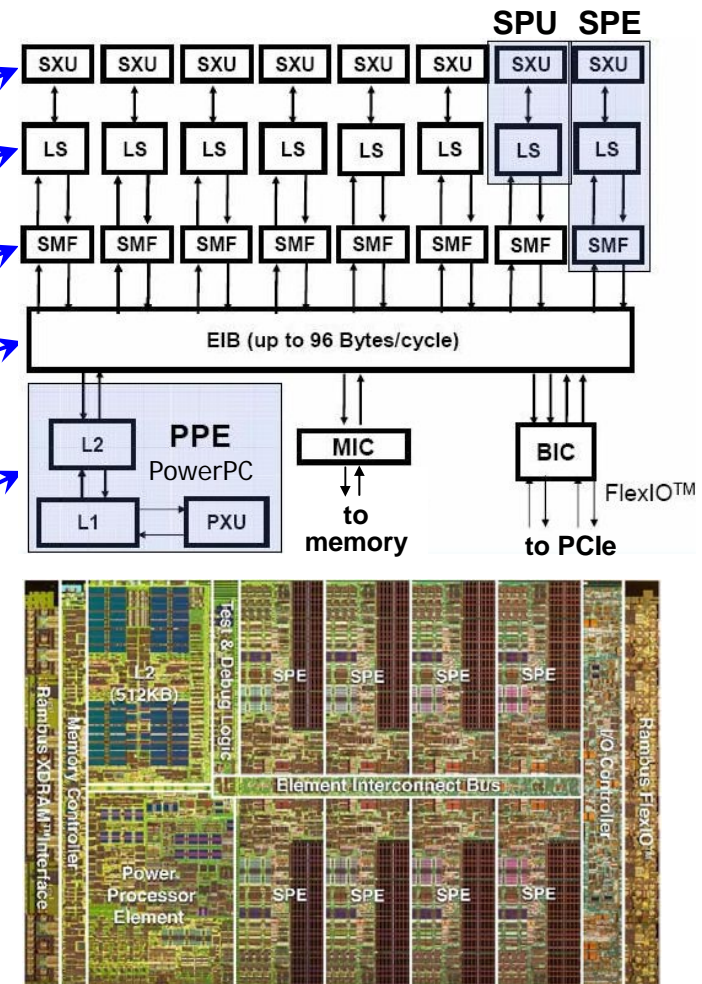
# Breakdown of Roadrunner configuration and speeds & feeds

---

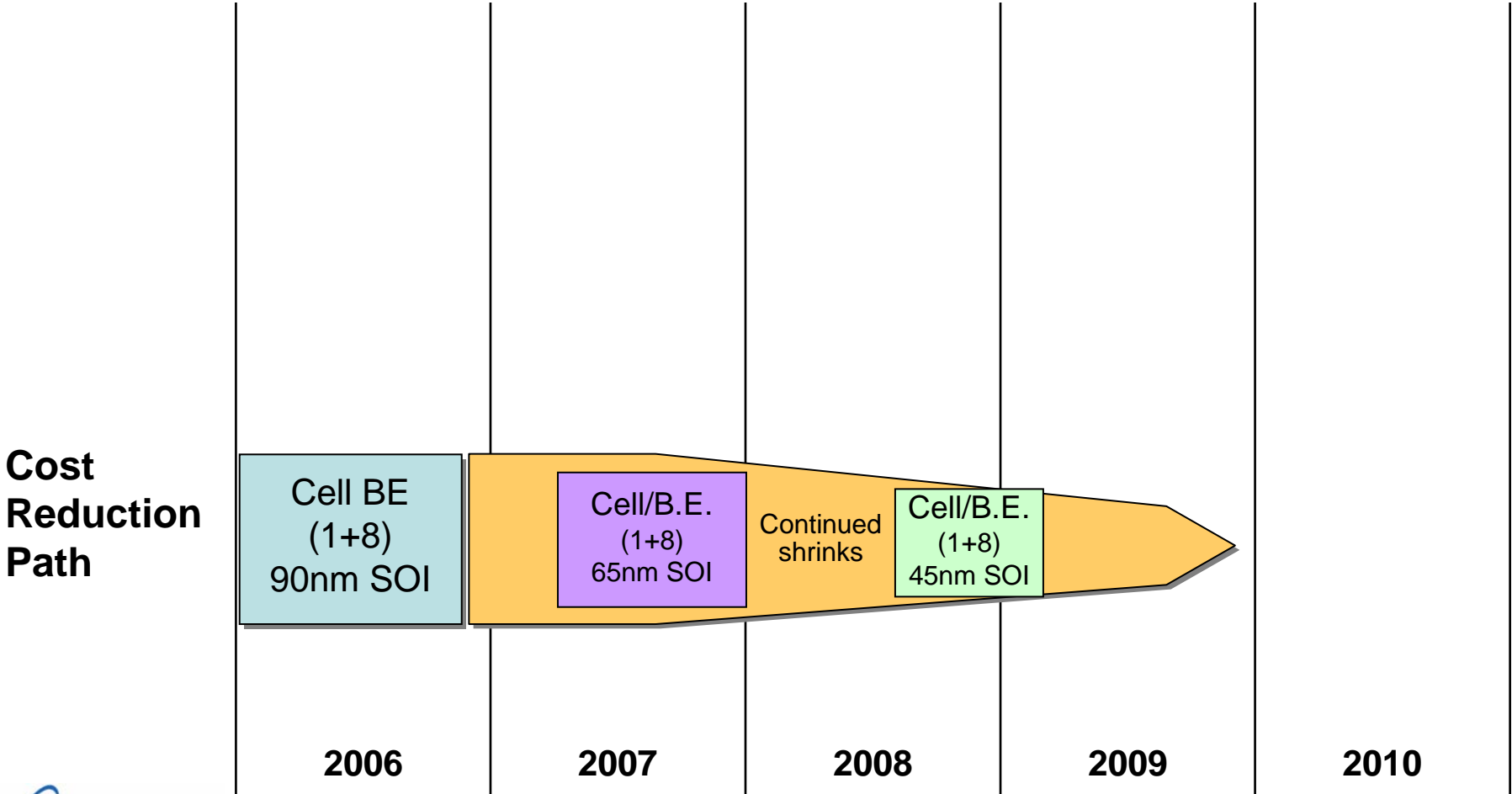
- Next few slides cover Roadrunner in this order:
  - *Cell processor*
  - *Accelerated Compute Node*
  - *Connected Unit cluster*
  - *Full system*
  - *Early-access prototype HW*

# The Cell processor is an (8+1)-way heterogeneous parallel processor

- Cell Broadband Engine (CBE\*) developed by Sony-Toshiba-IBM
  - used in Sony PlayStation 3
- 8 Synergistic Processing Elements (SPEs)
  - 128-bit vector engines
  - 256 kB local memory (LS = Local Store)
  - Direct Memory Access (DMA) engine (25.6 GB/s)
  - Chip interconnect (EIB)
  - Run SPE-code as POSIX threads (SPMD, MPMD, streaming)
- PowerPC PPE runs Linux OS
- Current performance:
  - 204.8 GF/s SP & 13.65 GF/s DP
  - 512 MB @ 25.6 GB/s XDR memory

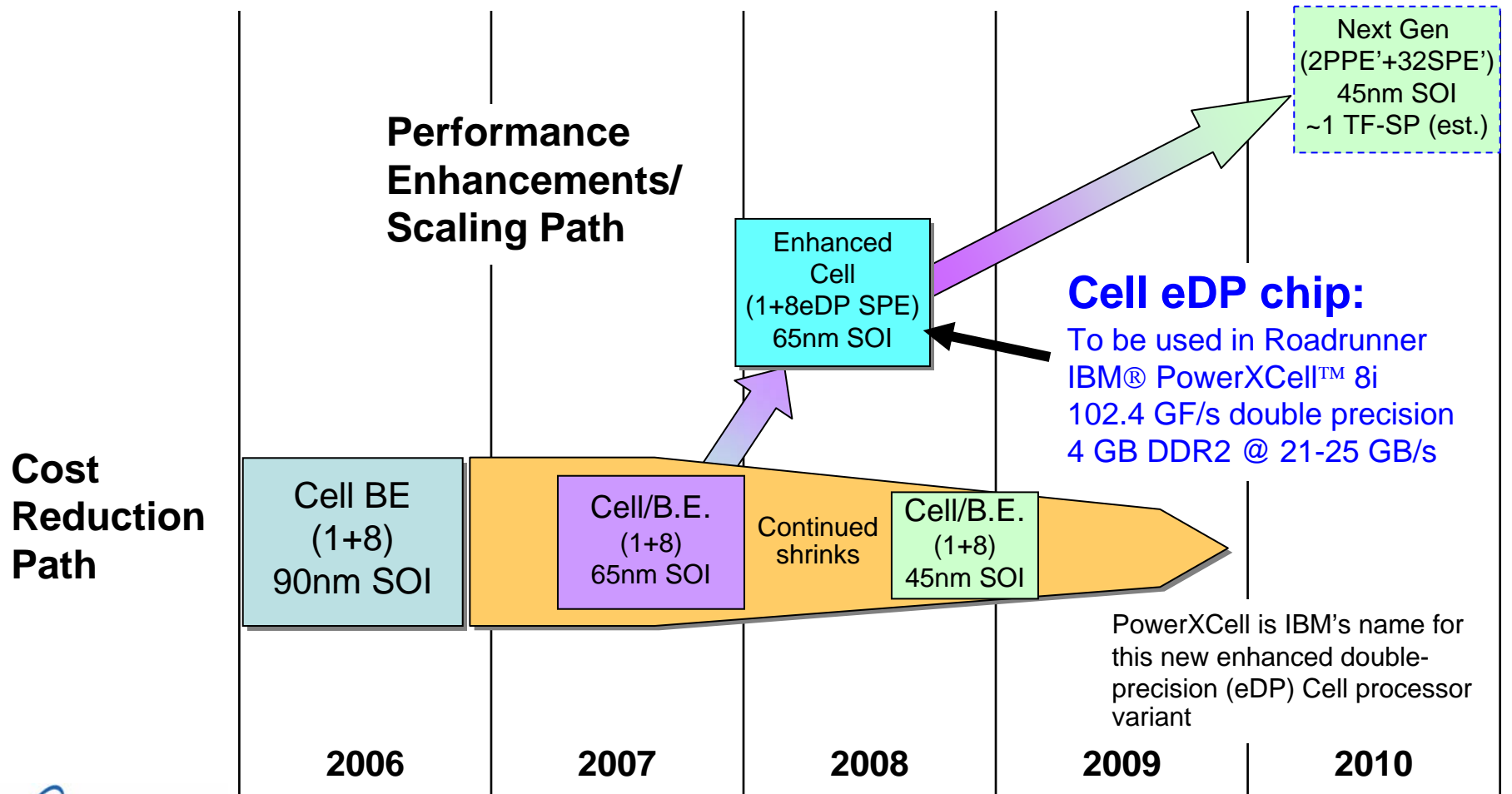


# Cell Broadband Engine™ Architecture (CBEA) Technology Competitive Roadmap



*All future dates and specifications are estimations only; Subject to change without notice. Dashed outlines indicate concept designs.*

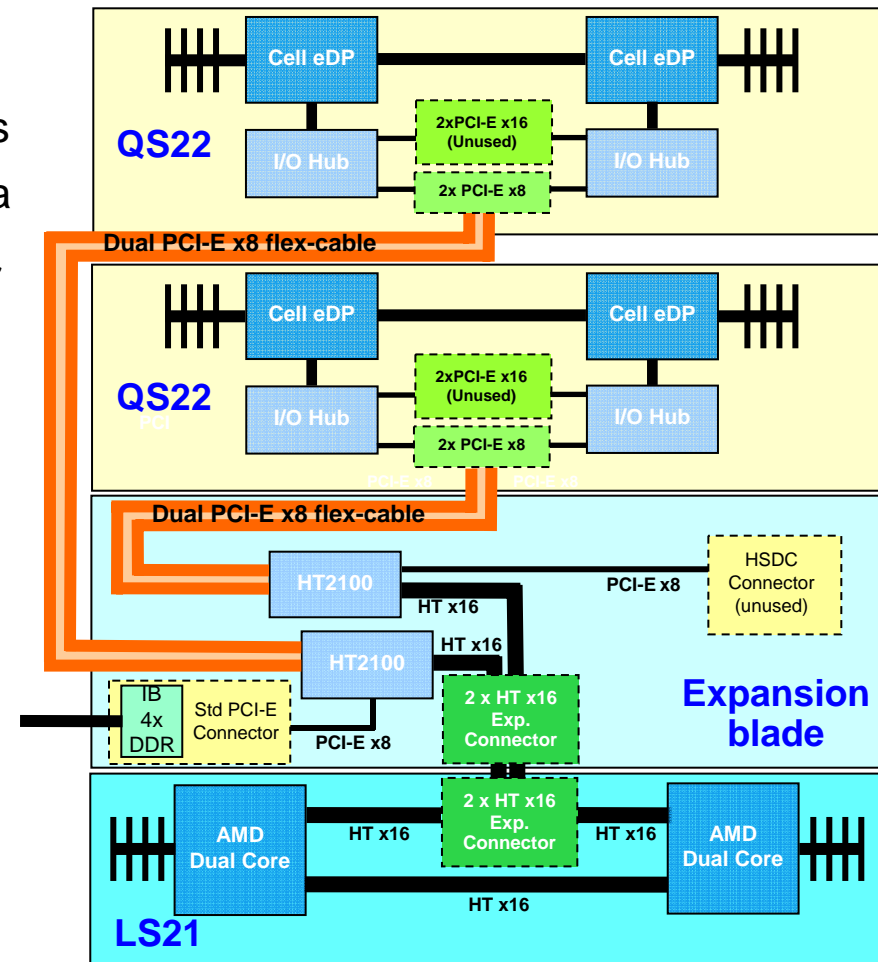
# Cell Broadband Engine™ Architecture (CBEA) Technology Competitive Roadmap



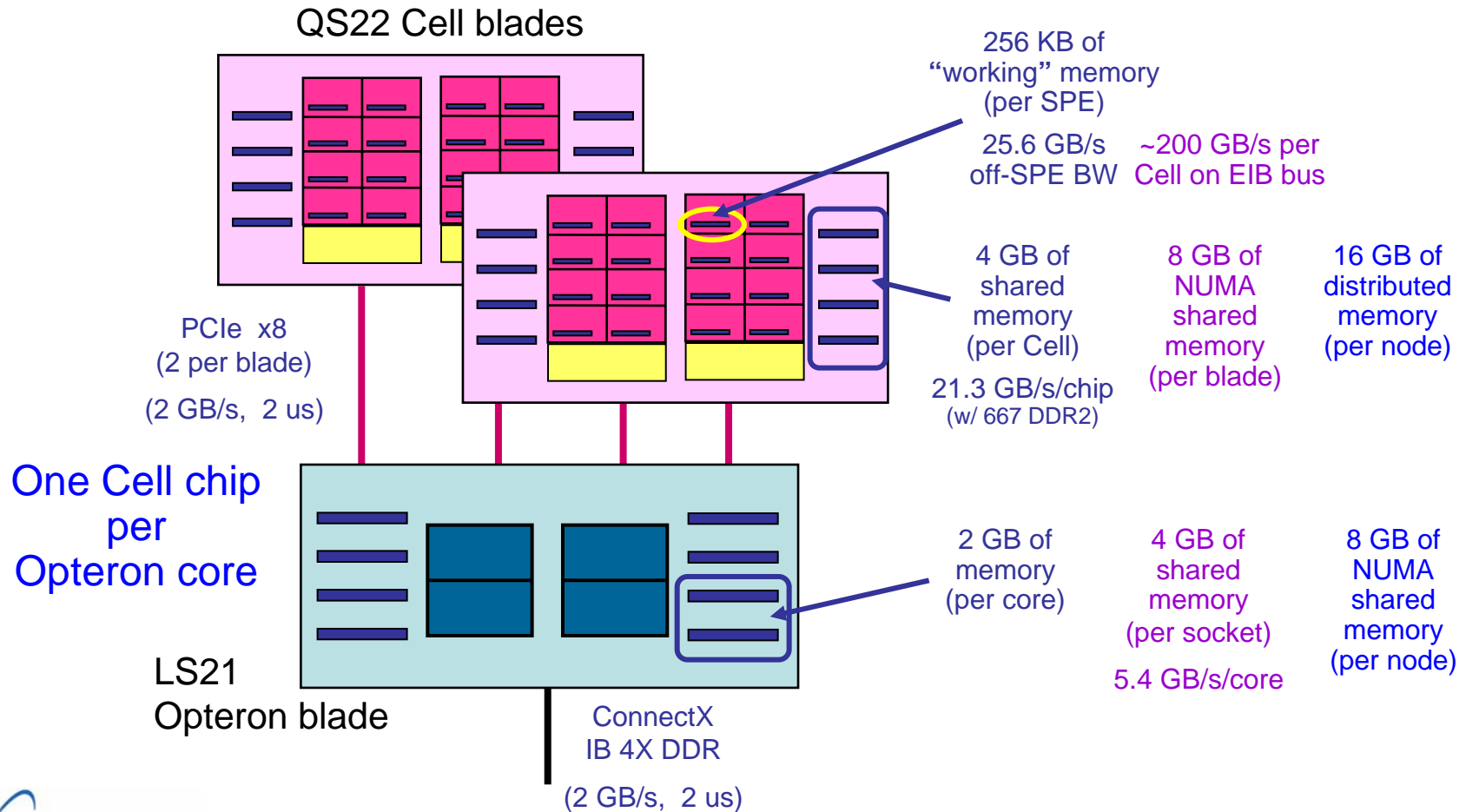
*All future dates and specifications are estimations only; Subject to change without notice. Dashed outlines indicate concept designs.*

# A Roadrunner Triblade node integrates Cell and Opteron blades

- **QS22** is a future IBM Cell blade containing two new enhanced double-precision (eDP/PowerXCell™) Cell chips
- Expansion blade connects two **QS22** via four **PCI-e x8** links to **LS21** & provides the node's ConnectX IB 4X DDR cluster attachment
- **LS21** is an IBM dual-socket Opteron blade
- 4-wide IBM BladeCenter packaging
- Roadrunner Triblades are completely diskless and run from RAM disks with NFS & Panasas only to the LS21
- Node design points:
  - *One Cell chip per Opteron core*
  - *~400 GF/s double-precision & ~800 GF/s single-precision*
  - *16 GB Cell memory & 8 GB Opteron memory*



# Roadrunner nodes have a memory hierarchy



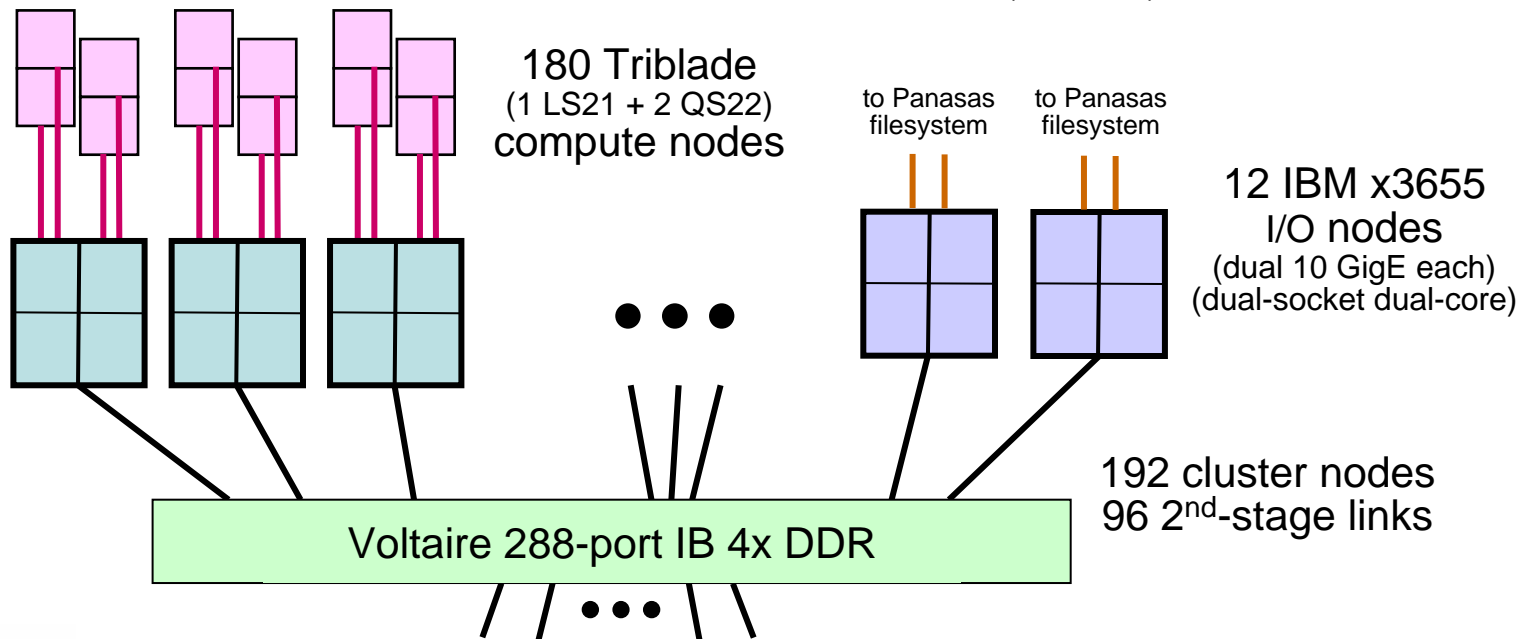
# A Connected Unit is a powerful cluster

## Connected Unit Specifications:

384 1.8 GHz dual-core Optrons  
 2.8 TF DP peak Optron  
 1.5 TB Optron memory

720 3.2 GHz Cell eDP chips  
 73.7 TF DP peak Cell eDP  
 2.88 TB Cell memory  
 15.4 TB/s Cell memory BW

192 IB 4X DDR cluster links  
 768 GB/s aggregate BW (bi-dir)  
 384 GB/s bi-section BW (bi-dir)  
 24 10 GigE I/O links on 12 I/O nodes  
 24 GB/s aggregate I/O BW (uni-dir)  
 (IB limited)



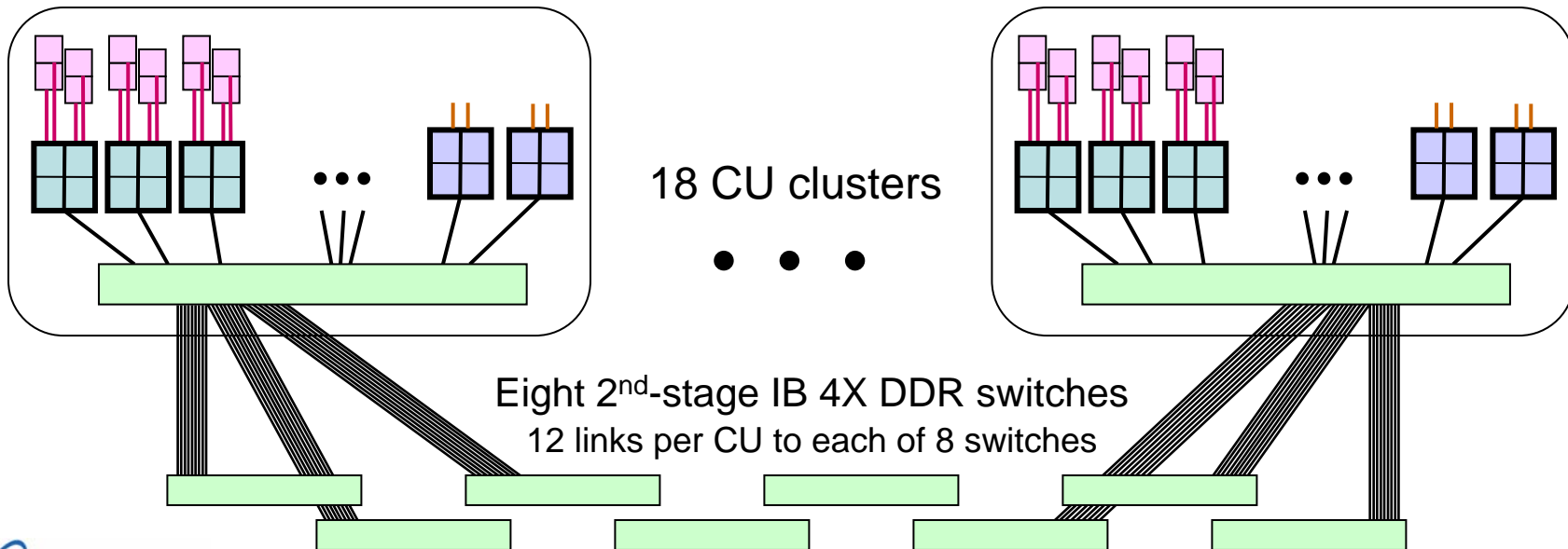
# Roadrunner is a petascale system in 2008

## Full Roadrunner Specifications:

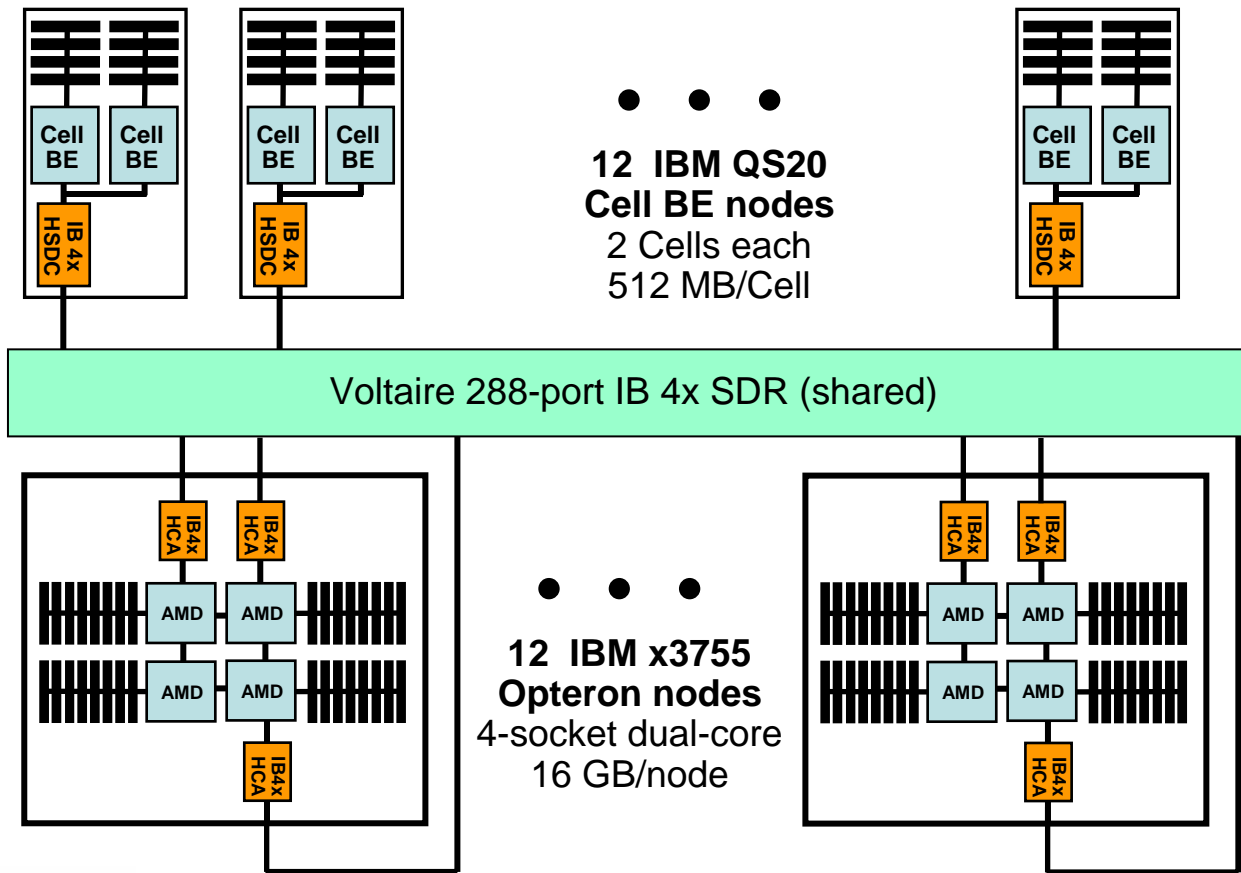
6,912 dual-core Opteron  
 49.8 TF DP peak Opteron  
 27.6 TB Opteron memory

12,960 Cell eDP chips  
 aka IBM PowerXCell™  
 1.33 PF DP peak Cell eDP  
 2.65 PF SP peak Cell eDP  
 51.8 TB Cell memory  
 277 TB/s Cell memory BW

3,456 nodes on 2-stage IB 4X DDR  
 13.8 TB/s aggregate BW (bi-dir) (1<sup>st</sup> stage)  
 6.9 TB/s aggregate BW (bi-dir) (2<sup>nd</sup> stage)  
 3.5 TB/s bi-section BW (bi-dir) (2<sup>nd</sup> stage)  
 432 10 GigE I/O links on 216 I/O nodes  
 432 GB/s aggregate I/O BW (uni-dir)  
 (IB limited)



# Prototype HW for applications testing



Phase 2  
prototype

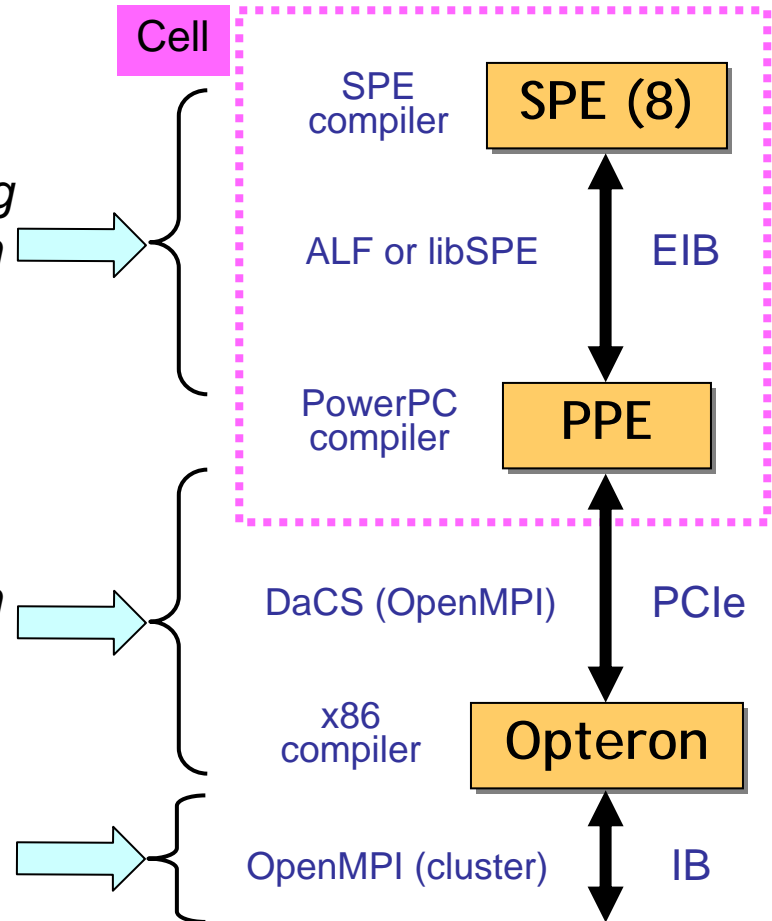
Advanced  
Architecture  
Initial  
System

aka. AAIS

(Operational  
January 2007)

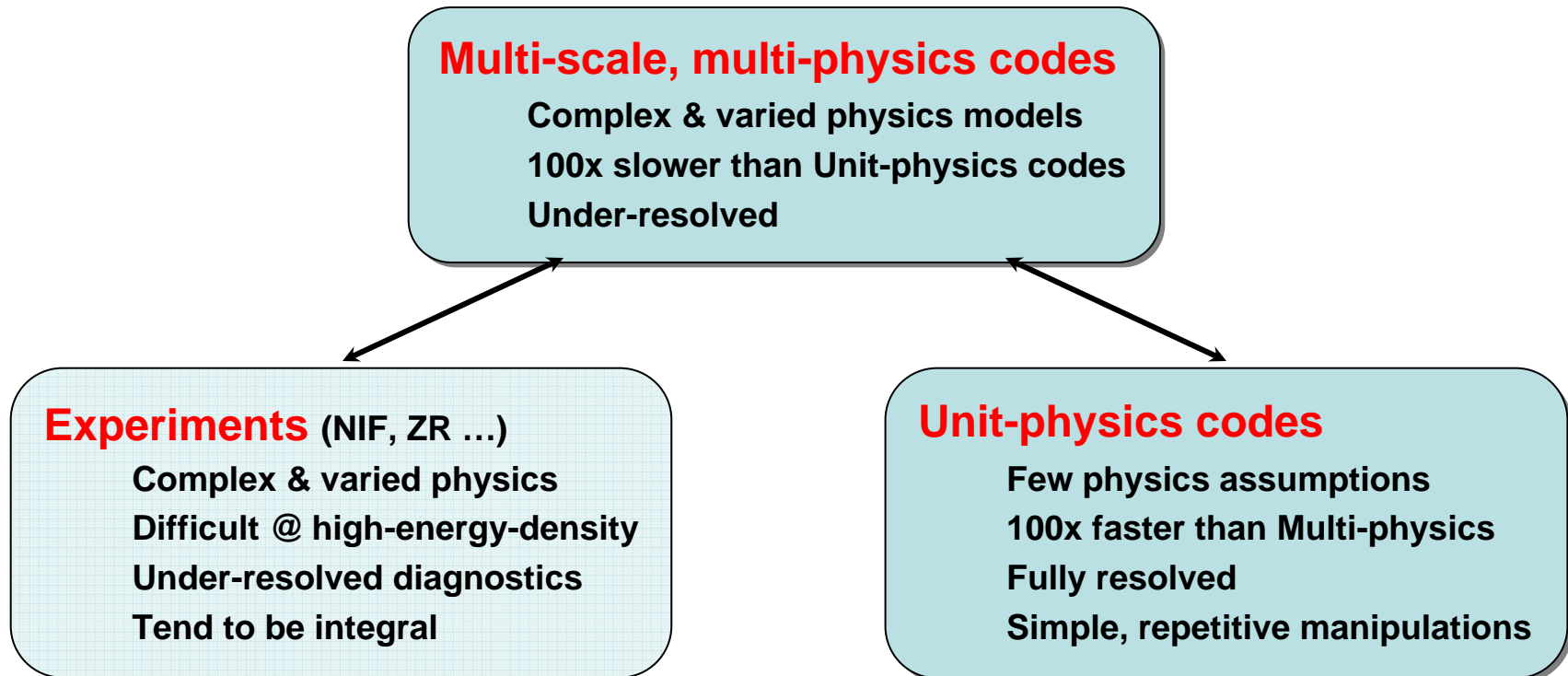
# Three types of processors work together.

- parallel computing on Cell
  - *data partitioning & work queue pipelining*
  - *process management & synchronization*
  
- remote communication to/from Cell
  - *data communication & synchronization*
  - *process management & synchronization*
  - *computationally-intense offload*
  
- **MPI remains as the foundation**

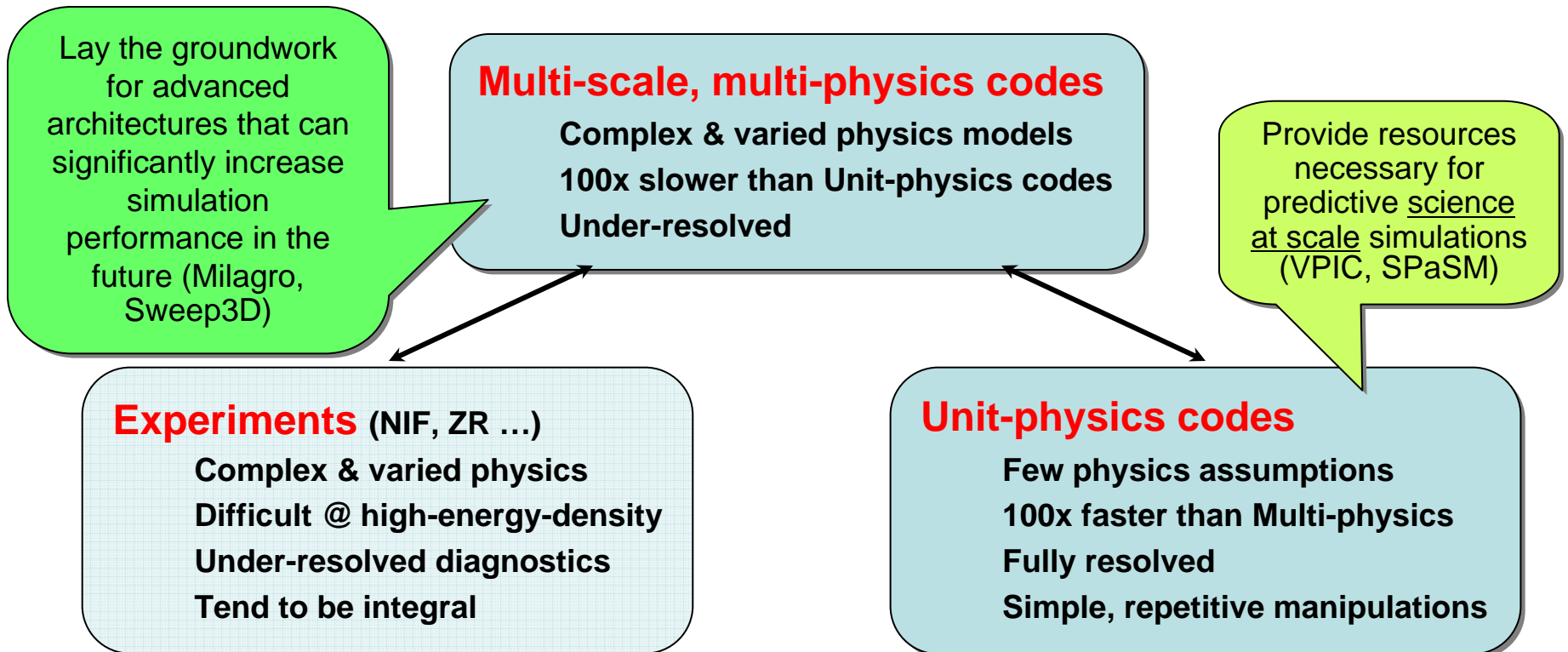


# Our vision for high-performance computing embraces both multi- and unit-physics codes

---



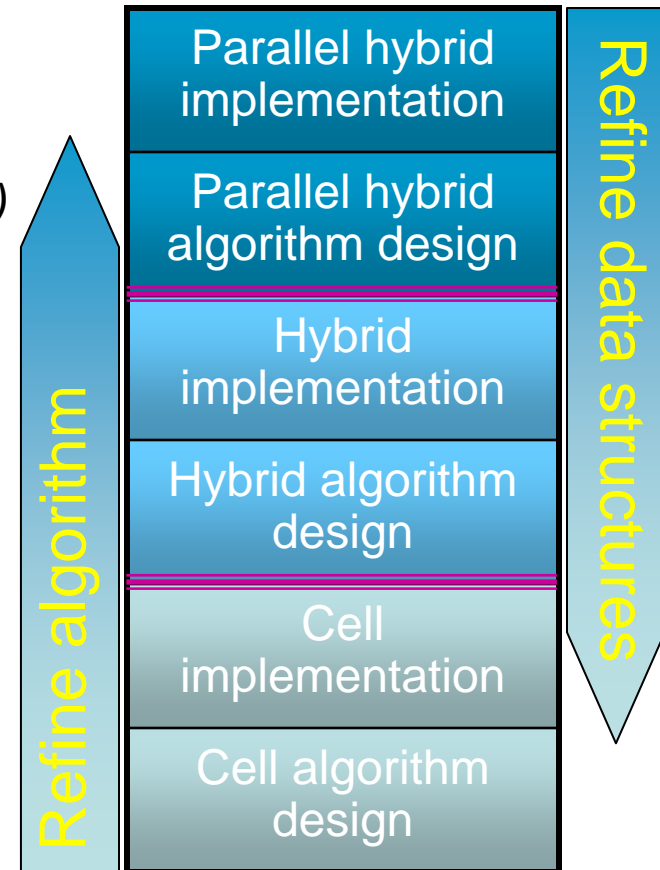
# Our vision for high-performance computing embraces both multi- and unit-physics codes



**Roadrunner is about the future.**

# A few important key algorithms are being targeted

- Transport
  - *PARTISN (Sn neutron transport)*
    - Sweep3D (benchmark code)
    - Sparse solver (PCG)
  - *MILAGRO (IMC thermal radiation transport)*
- Particle methods
  - *VPIC (Particle-In-Cell)*
    - SSE enabled
  - *SPaSM (molecular dynamics)*
    - Data parallel CM-5 implementation
- Eulerian hydro
  - *Direct Numerical Simulation*
- Linear algebra
  - *LINPACK*
  - *Preconditioned Conjugate Gradient (PCG)*



# Roadrunner at a glance

- **Cluster of 18 Connected Units (CU)**
  - 6,912 AMD dual-core Opterons
  - 12,960 IBM Cell eDP accelerators
  - 49.8 Teraflops peak (Opteron)
  - 1.33 Petaflops peak (Cell eDP)
  - 1PF sustained Linpack
- **InfiniBand 4x DDR fabric**
  - 2-stage fat-tree; all-optical cables
  - Full bi-section BW within each CU
    - 384 GB/s (bi-directional)
  - Half bi-section BW among CUs
    - 3.45 TB/s (bi-directional)
  - Non-disruptive expansion to 24 CUs
- **80 TB aggregate memory**
  - 28 TB Opteron
  - 52 TB Cell
- **216 GB/s sustained File System I/O:**
  - 216x2 10G Ethernets to Panasas
- **RHEL & Fedora Linux**
- **SDK for Multicore Acceleration**
  - Cell compilers, libraries, tools
- **xCAT Cluster Management**
  - System-wide GigE network
- **3.9 MW Power:**
  - 0.35 GF/Watt
- **Area:**
  - 296 racks
  - 5500 ft<sup>2</sup>



The End



Operated by the Los Alamos National Security, LLC for the DOE/NNSA

