

UNCLASSIFIED

**WEAPONS** *SCIENCE & ENGINEERING*  
*CAPABILITY REVIEW*

# Roadrunner and hybrid computing

Ken Koch

Roadrunner Technical Manager  
Scientific Advisor, CCS-DO

August 22, 2007



LA-UR-07-6213

Operated by Los Alamos National Security, LLC for NNSA

UNCLASSIFIED



# Outline

---

1. Roadrunner goals and thrusts
2. Hybrid computing and the Cell processor
3. Roadrunner system and Cell-acceleration architecture
4. Overview of programming concepts
5. Overview of algorithms & applications

UNCLASSIFIED

**WEAPONS** *SCIENCE & ENGINEERING*  
*CAPABILITY REVIEW*

**Roadrunner is essential for fulfilling our  
science and stewardship missions for  
the weapons program**

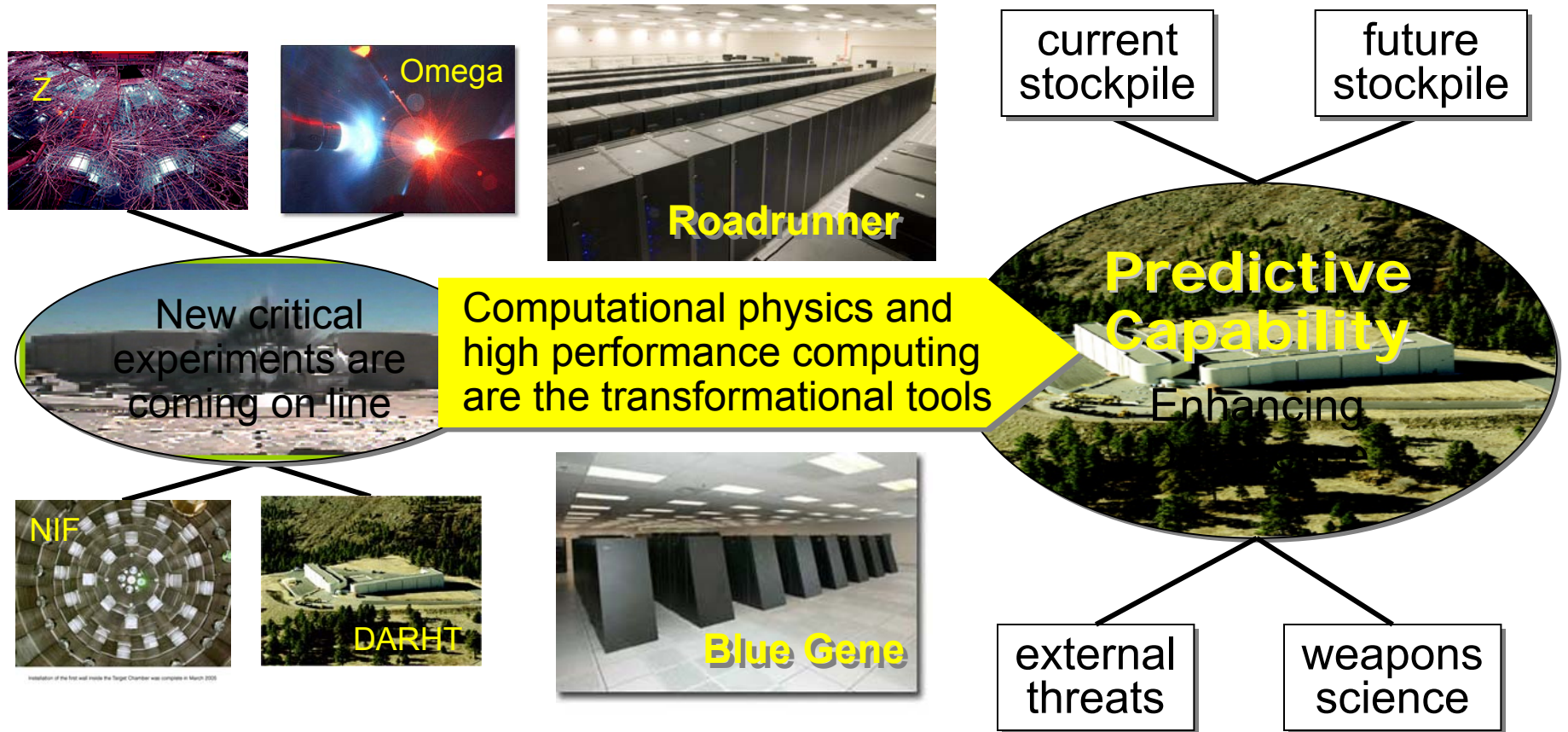


Operated by Los Alamos National Security, LLC for NNSA

UNCLASSIFIED



# Petascale computing is essential to ensuring a sustainable nuclear deterrence



# The Cell-accelerated Roadrunner targets two goals

---

- Science @ Scale
  - Multi-scale unit physics for weapons & open science
    - Validate model assumptions
    - Better understand physics
    - Cross-validate physics models at overlapping resolutions
  - Run at Petascale (25% to 80+% of machine)
- Advanced Architecture for algorithms and applications
  - Target select physics and work on algorithms and implementations
    - Convert algorithms to Cell & Roadrunner
    - Or use an alternative or modified parallel algorithm
  - Provide faster solutions or improved accuracy
  - Incrementally update existing ASC integrated codes for targeting key uncertainties
    - Target focused simulations, not general usage
    - Focus is more predictive science oriented than speeding up production jobs
- *More on this at end of talk*

UNCLASSIFIED

# Heterogeneous & hybrid computing is an industry trend

## The Cell processor is a powerful hybrid processor that can accelerate algorithms



Operated by Los Alamos National Security, LLC for NNSA

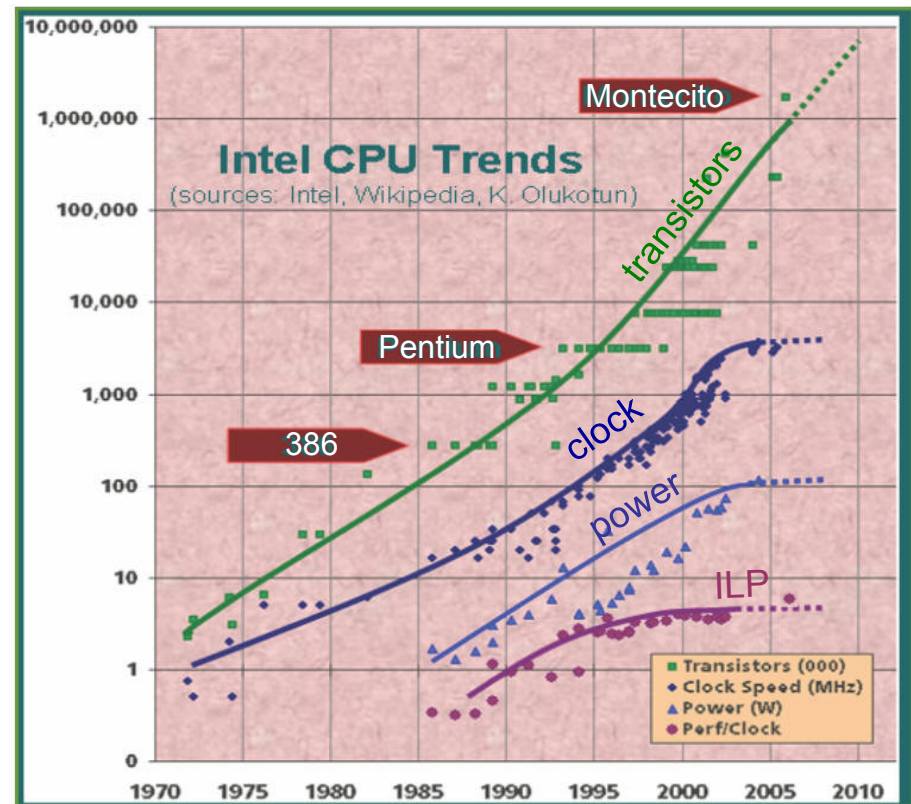
**WEAPONS SCIENCE & ENGINEERING  
CAPABILITY REVIEW**



UNCLASSIFIED

## Microprocessor trends are changing

- Moore's law still holds, but is now being realized differently
  - Frequency, power, & instruction-level-parallelism (ILP) have all plateaued
  - Multi-core is here today and many-core ( $\geq 32$ ) looks to be the future
  - Memory bandwidth and capacity per core are headed downward (caused by increased core counts)
  - Key findings of Jan. 2007 IDC Study: "Next Phase in HPC"
    - *new ways of dealing with parallelism will be required*
    - *must focus more heavily on bandwidth (flow of data) and less on processor*

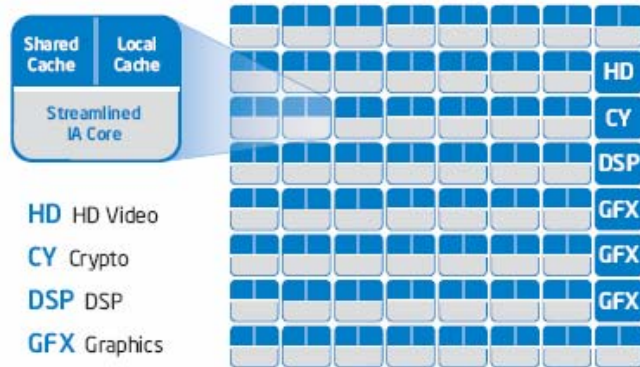


From Burton Smith, LASCI-06 keynote, with permission

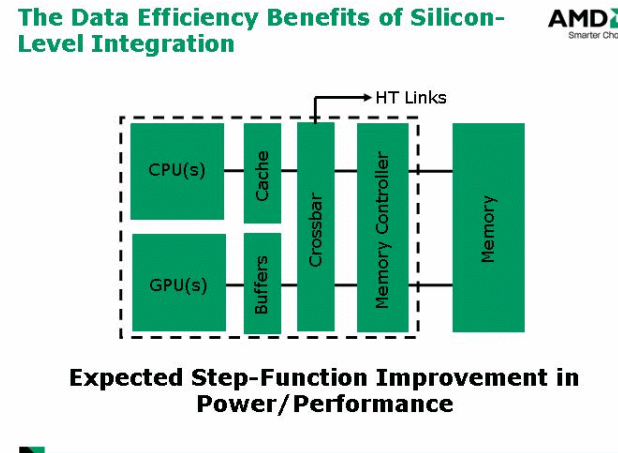
**WEAPONS SCIENCE & ENGINEERING  
CAPABILITY REVIEW**

# Industry presentations show changing trends in processors

## Intel's Microprocessor Research Lab



## AMD Fusion



## Intel's Visual Computing Group - Larabee



October 2006 Unleashing the Processing Powerhouse

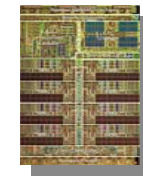
## nVidia G80 - 2006





# The change is already happening

- **New processors & accelerators:**
  - Multi-core to many-core: 8, 16, 32, ... 80, 128, 1000?
  - IBM Cell , AMD Fusion, Intel Polaris, NVidia G8800
  - Distributed memory & caches at core level
  - FPGAs, GPGPU, Clearspeed CSX600, IBM Cell, XtremeData XD1000, Nvidia G80, AMD Stream Processor
- **Connection standards**
  - AMD Torrenza, Intel/IBM Geneseo, AMD HyperTransport Initiative
- **Programming**
  - IBM Roadrunner Cell libraries, RapidMind, Peakstream, Impulse C, Stanford's Sequoia, NVidia CUDA, Clearspeed C, Mercury MFC, stream programming
- **Heterogeneous architectures**
  - Clusters of mixed node
  - Hybrid accelerated node (e.g. [Roadrunner](#), Clearspeed, FPGA)
  - Hybrid on the same bus (e.g. CPU+GPUs, Intel's Geneseo, AMD's Torrenza)
  - Within processor itself (e.g. Cell, AMD Fusion)



Cell or  
DSP or  
FPGA

+



CPU

+



GPU

=

hybrid

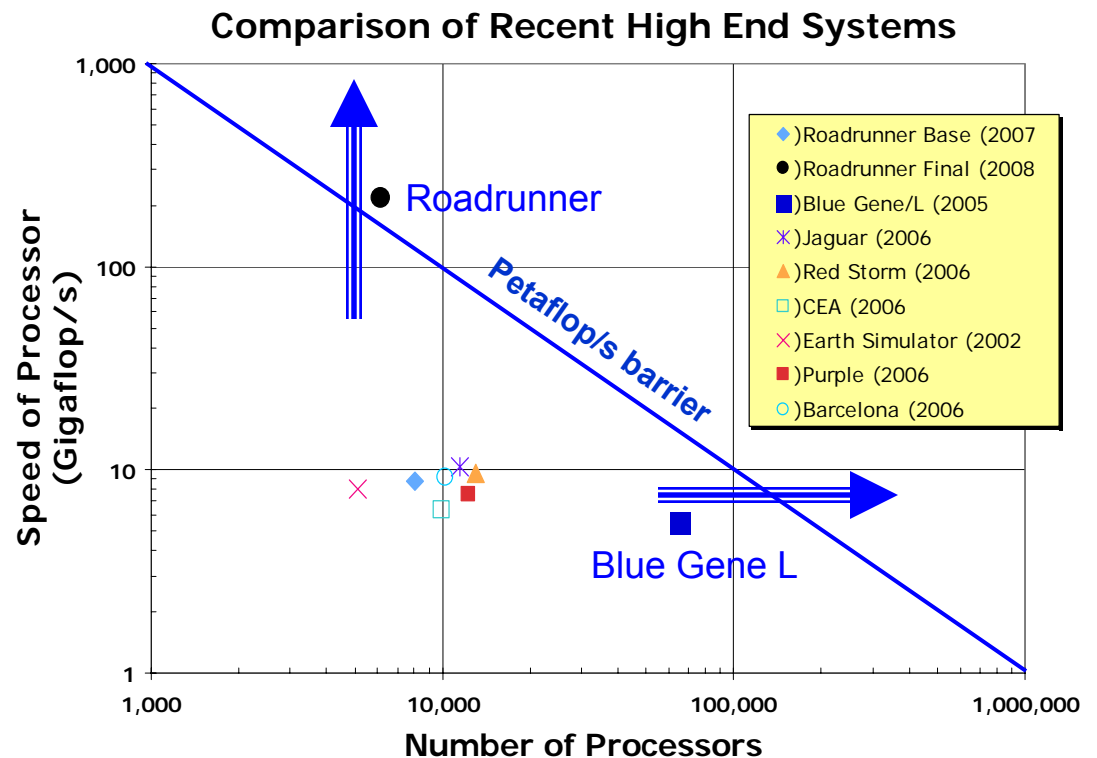
## There were hybrid computers before Roadrunner

---

- Floating Point Systems FPS Array Processors (AP-120B, FPS-164/264) (circa 1976-1982)
  - [http://en.wikipedia.org/wiki/Floating\\_Point\\_Systems](http://en.wikipedia.org/wiki/Floating_Point_Systems)
- Deep Blue for chess (IBM SP-2: 30 RS6K + 480 chess chips) (circa 1997)
  - [http://en.wikipedia.org/wiki/Deep\\_Blue](http://en.wikipedia.org/wiki/Deep_Blue)
- Grape-6 for stellar dynamics w/ custom chips) (circa 2000-2004)
  - <http://grape.astron.s.u-tokyo.ac.jp/~makino/grape6.html>
- Various FPGA supercomputers from system vendors:
  - SRC-6 (w/ MAP), Cray XD1 (w/ Application Acceleration), SGI Altix (w/ RASC)
- Titech TSUBAME (w/ some Clearspeed) (2006)
  - <http://www.gsic.titech.ac.jp/English/Publication/pressrelease.html.en>
- RIKEN MDGrape-3 “Protein Explorer” (w/ custom chips) (2006)
  - <http://mdgrape.gsc.riken.jp/modules/tinyd0/index.php>
- Terra Soft’s Cell E.coli/Amoeba PS3 Cluster (cluster of 1U PlayStation 3 development systems) (2007)
  - <http://www.hpcwire.com/hpc/967146.html>

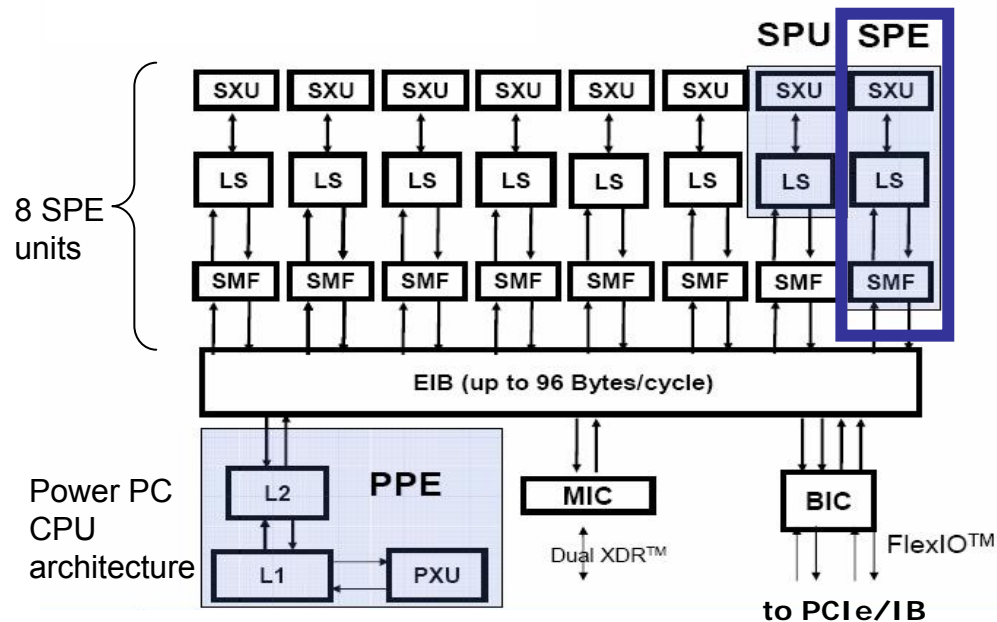
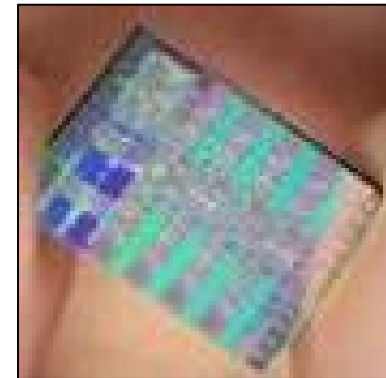
# Roadrunner is paving the way along an alternate path for the future of HPC

- Roadrunner is the first of a new breed of high performance computers
- Roadrunner is sooner, cheaper, and smaller than building a petascale machine in the conventional way
- Roadrunner at Los Alamos attacks now the unavoidable software challenge early
  - “The Labs must be in the game now.” LANS *Independent Functional Management Assessment of RR project (May, 2007)*



# The Cell is a powerful hybrid multi-processor chip

- Cell Broadband Engine™ \* (Cell BE)
  - Developed under Sony-Toshiba-IBM efforts
  - Current Cell chip is used in the Sony PlayStation 3
- An 8-way heterogeneous parallel engine



Each of the 8 SPEs are 128 bit (e.g. 2-way DP-FP) vector engines w/ 256KB of Local Store (LS) memory & a DMA engine.

They can operate together or independently (SPMD or MPMD).

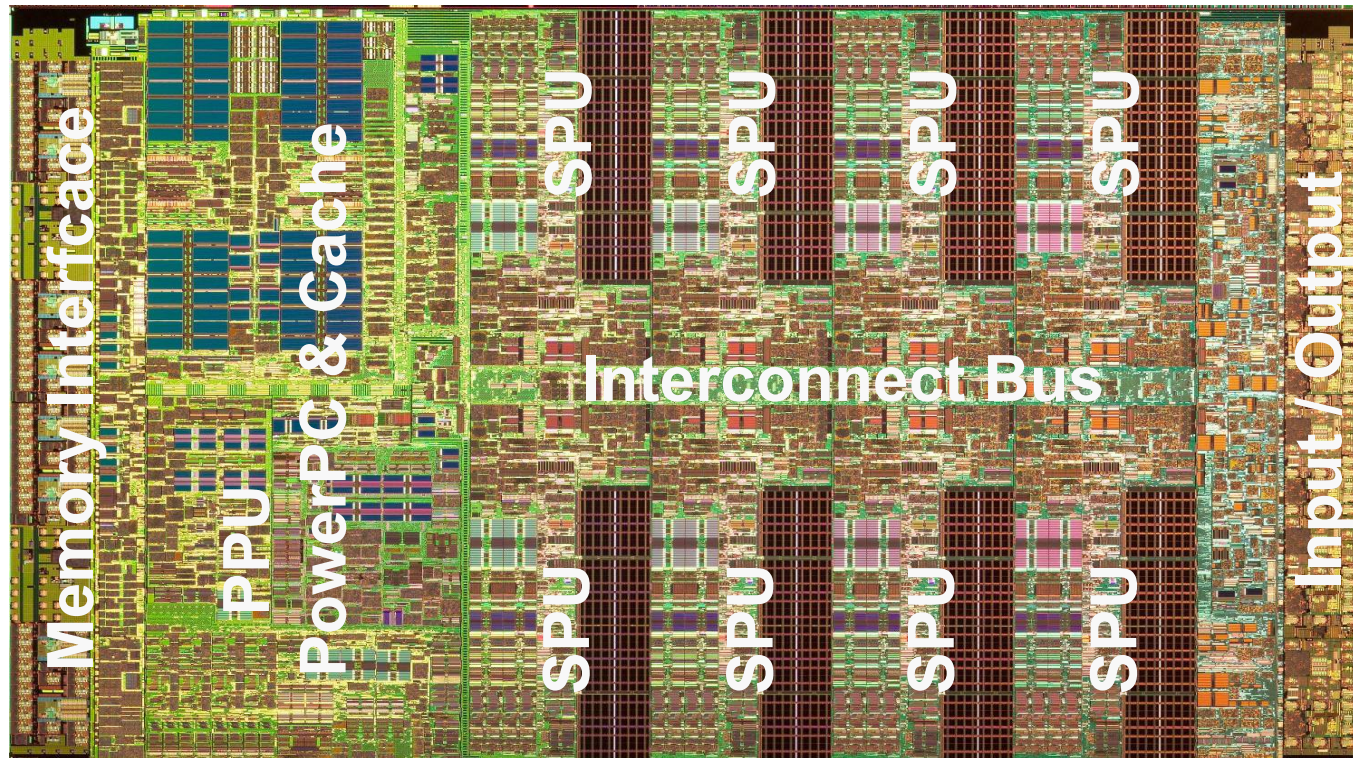
~200 GF/s single precision

~ 15 GF/s double precision (current chip)

\* Trademark of Sony Computer Entertainment, Inc.

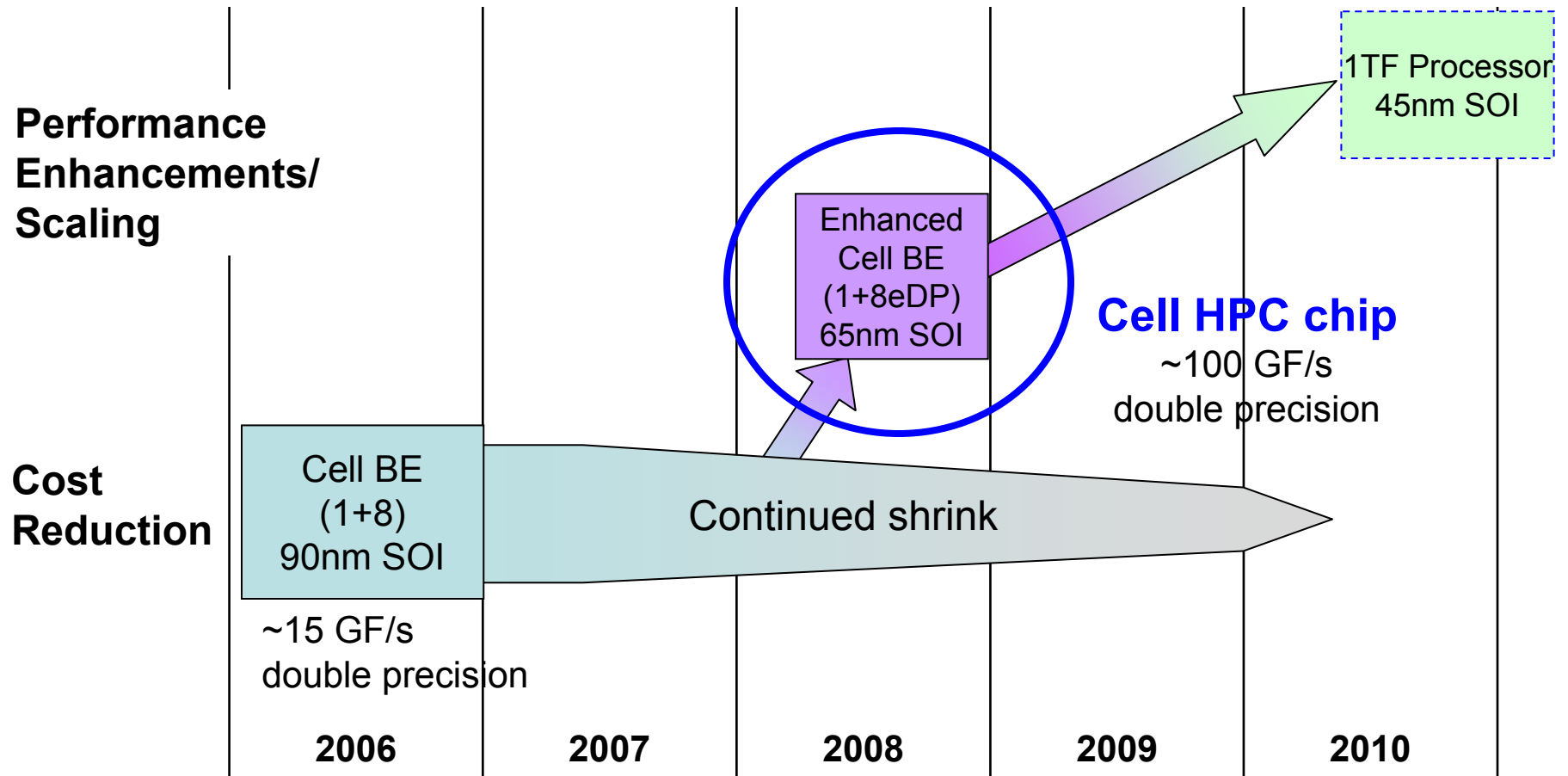
**WEAPONS SCIENCE & ENGINEERING  
CAPABILITY REVIEW**

# Cell Broadband Engine



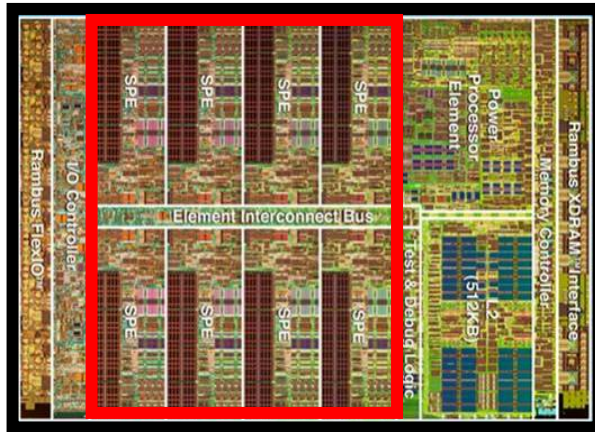
Heterogeneous: 1 PPU + 8 SPUs

# A new Cell rev. was needed for Roadrunner

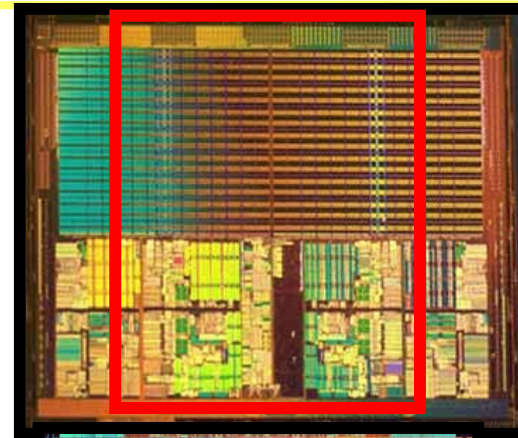


# Cell local store architecture has a performance per area advantage

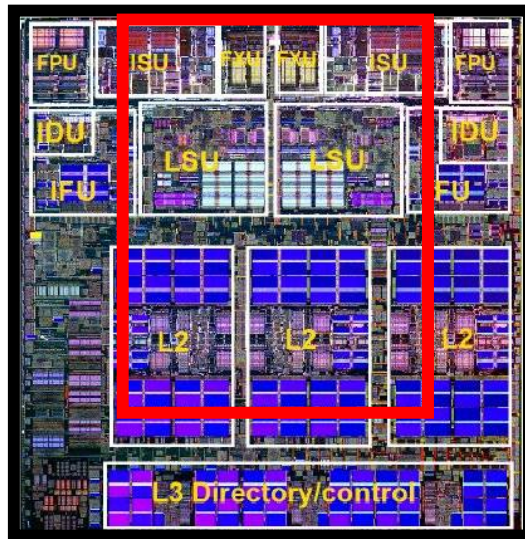
*Cell HPC*  
~100 GF/s



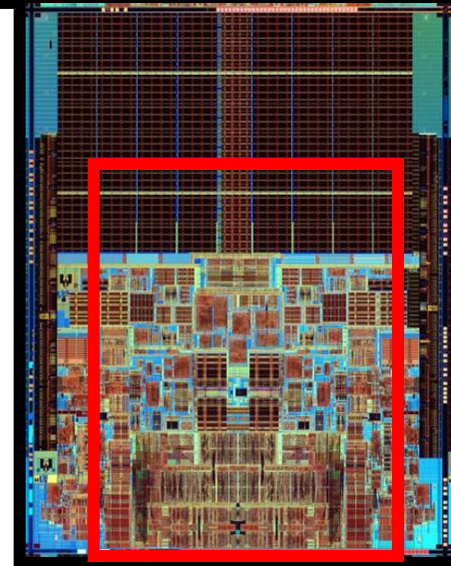
*AMD*  
~20 GF/s



*IBM*  
~20 GF/s

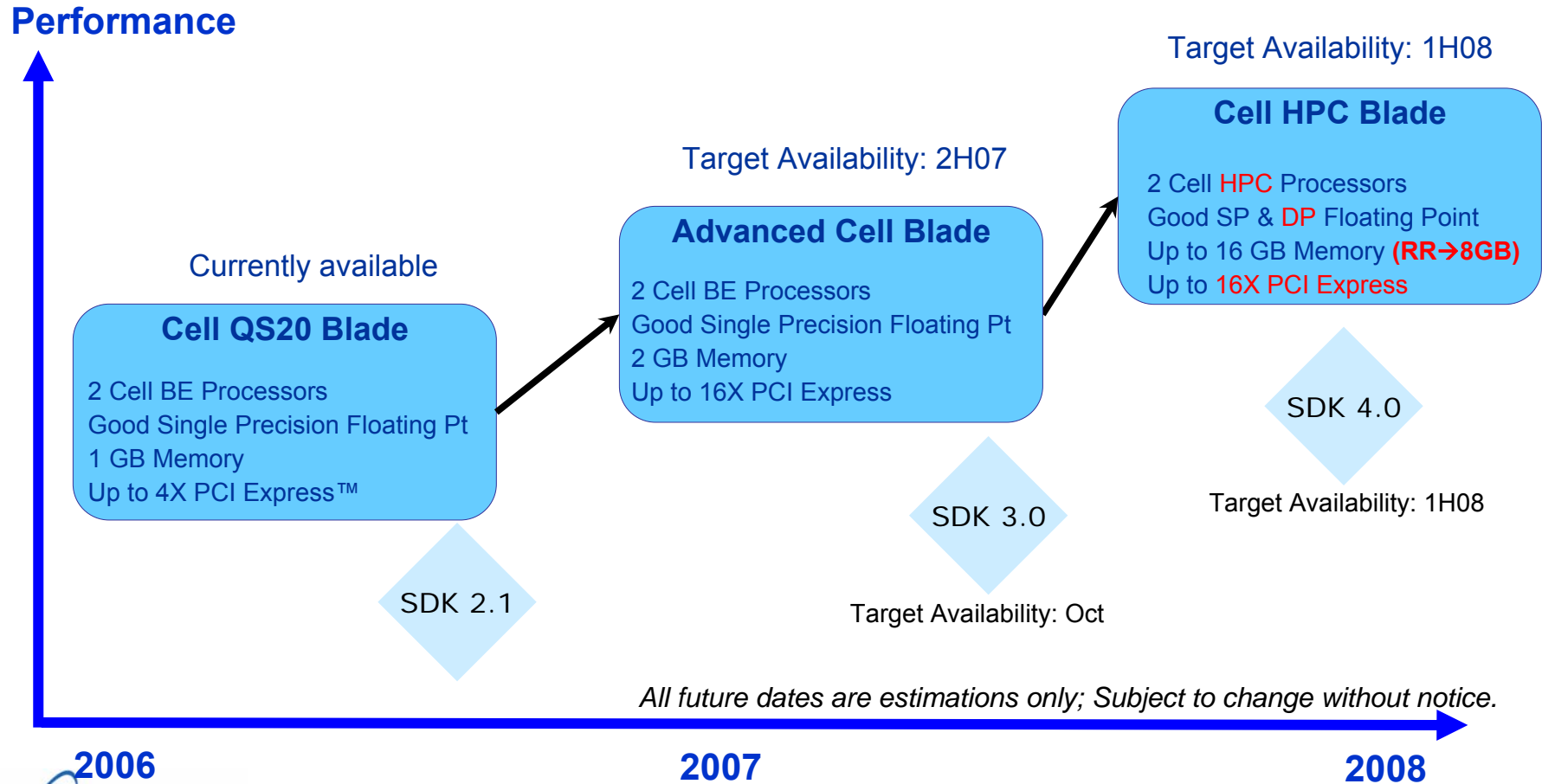


*Intel*  
~20 GF/s



**WEAPONS SCIENCE & ENGINEERING  
CAPABILITY REVIEW**

# Cell blades and software improve during Roadrunner development period





UNCLASSIFIED

**Roadrunner is a hybrid architecture  
for 2008 deployment achieving a  
sustained PetaFlop/s performance level**




Operated by Los Alamos National Security, LLC for NNSA

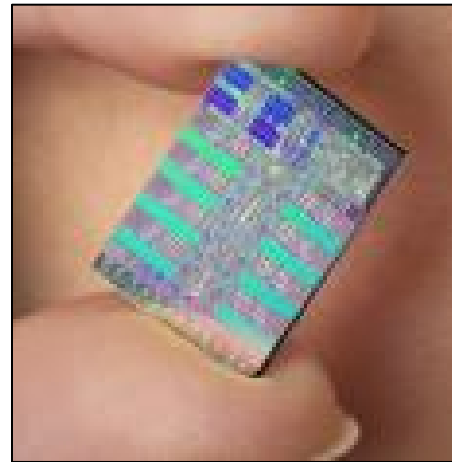
**WEAPONS SCIENCE & ENGINEERING  
CAPABILITY REVIEW**



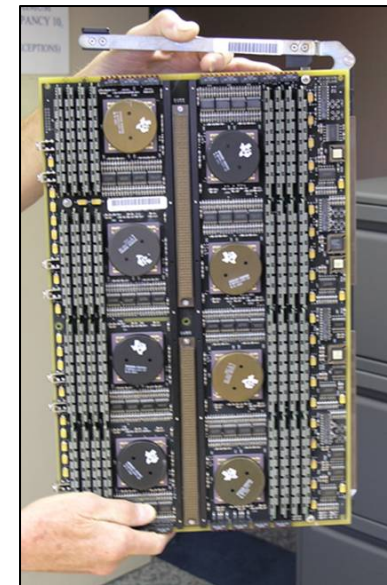
UNCLASSIFIED

## Roadrunner project is a partnership with IBM

- Contract signed September 8, 2006 with 
- Critical component of stockpile stewardship
  - **Phase 1** (Base system) supports near-term mission deliverables
  - **Phase 2** (Cell prototypes) supports pre-Final assessment
  - **Phase 3** (Hybrid final system)
    - Achieves PetaFlops level of performance
    - Demonstrates new paradigm for high performance computing
- Accelerated vision of the future



Cell processor (2007, 100 GF)



CM-5 board (1994, 1 GF)

100x in  
14 yrs  
8 vector units each



**WEAPONS SCIENCE & ENGINEERING  
CAPABILITY REVIEW**



# Status of the Roadrunner Phases

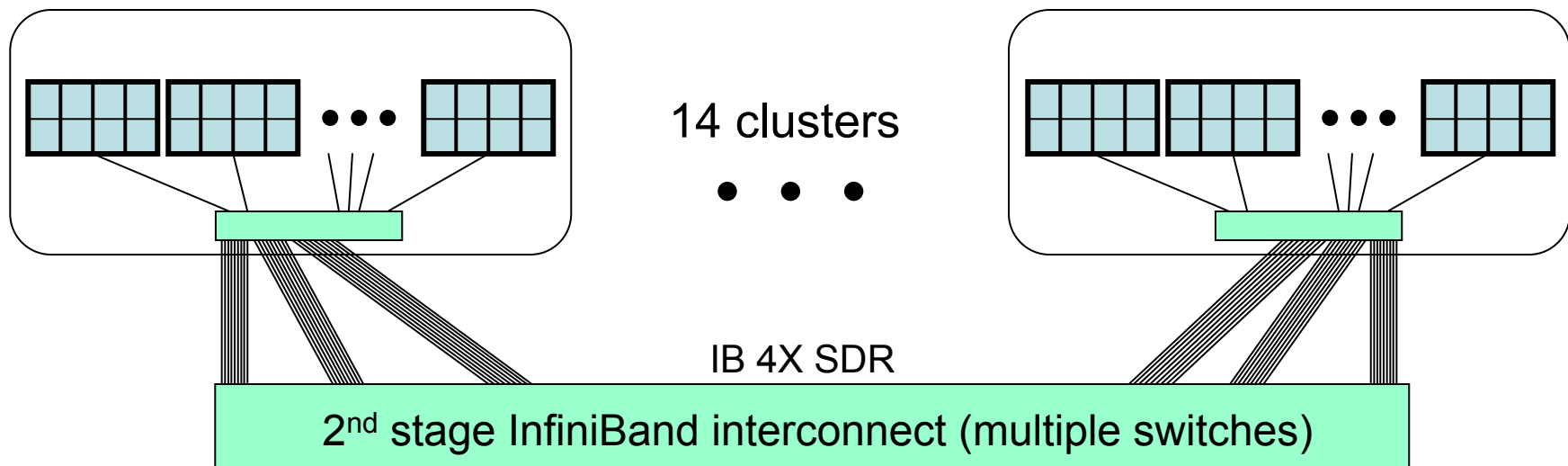
---

- **Phase 1**
  - Known as “**Base**” system
  - 14 clusters of Opteron-only nodes
  - In classified operation now with general availability early September
  - Already contributing to DSW efforts
    - Application Readiness Team (ART) provided early support
- **Phase 2**
  - Known as “**AAIS**” (Advanced Architecture Initial System)
  - 6 Opteron nodes and 14 Cell blades on Infiniband
  - Has been in use since January for Cell & hybrid application prototyping
- **Phase 3**
  - Contract option for 2008 delivery
  - Two technical Assessments scheduled for this October
  - A redesigned Cell-accelerated system for better performance
    - No longer an upgrade to the Base system

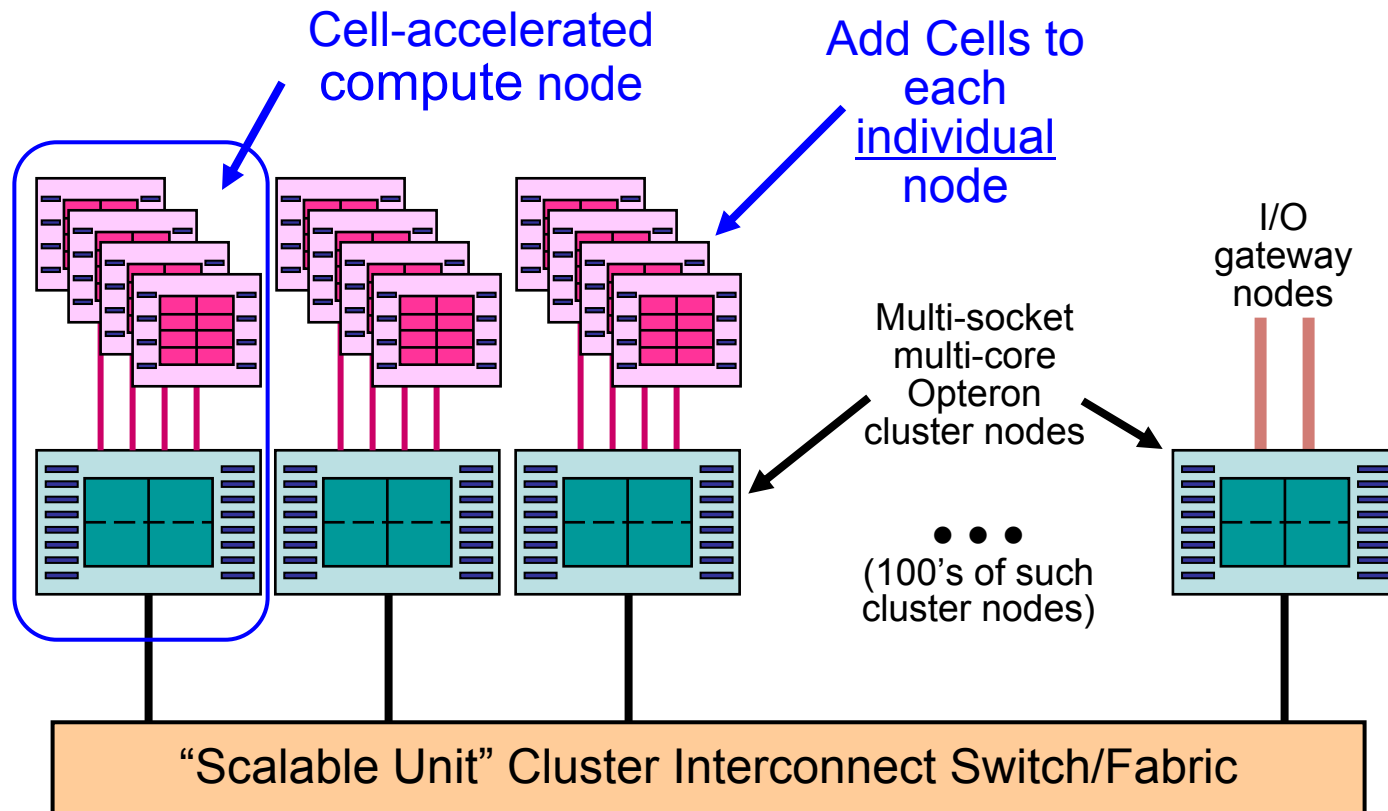
# Roadrunner Phase 1 “Base” is deployed now as a capacity resource



- Fourteen InfiniBand-interconnected clusters of Opteron nodes
  - Provides 70 Teraflop/s of needed capacity computing resources
  - In classified operation now with early September general availability
  - 144 4-socket dual-core 32 GB memory nodes per cluster



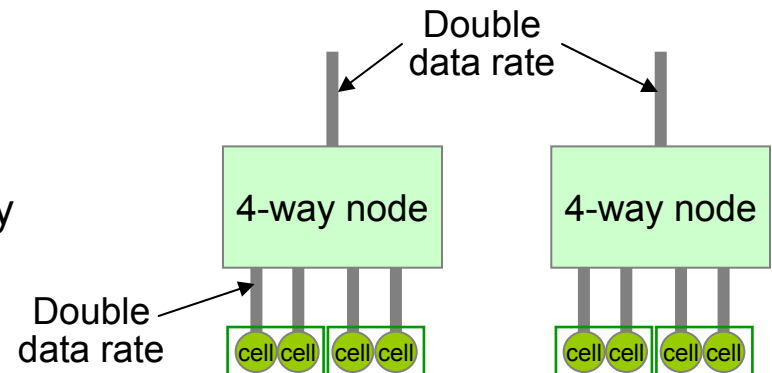
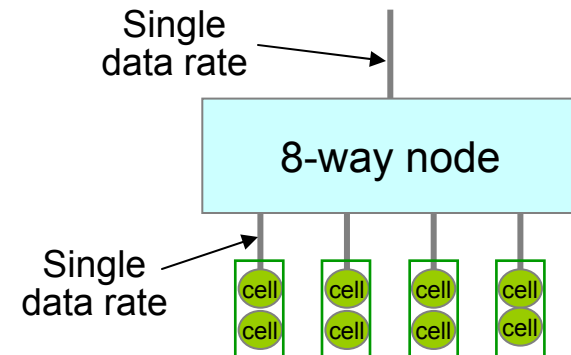
# Roadrunner Phase 3 is Cell-accelerated, not a cluster of Cells



This is what makes Roadrunner different!

## Phase 3 was redesigned to be a better system

- Phase 3 was to be an upgrade
  - Simply add InfiniBand connected Cell blades to Phase 1
- Phase 3 is now an entirely new additional system
  - Uses smaller node with PCI Express connected Cell blades
  - Better performance on same schedule
  - LANL keeps existing Base system
    - Classified capacity work not disturbed by a major upgrade
    - “Science” runs and presence in the open is now possible

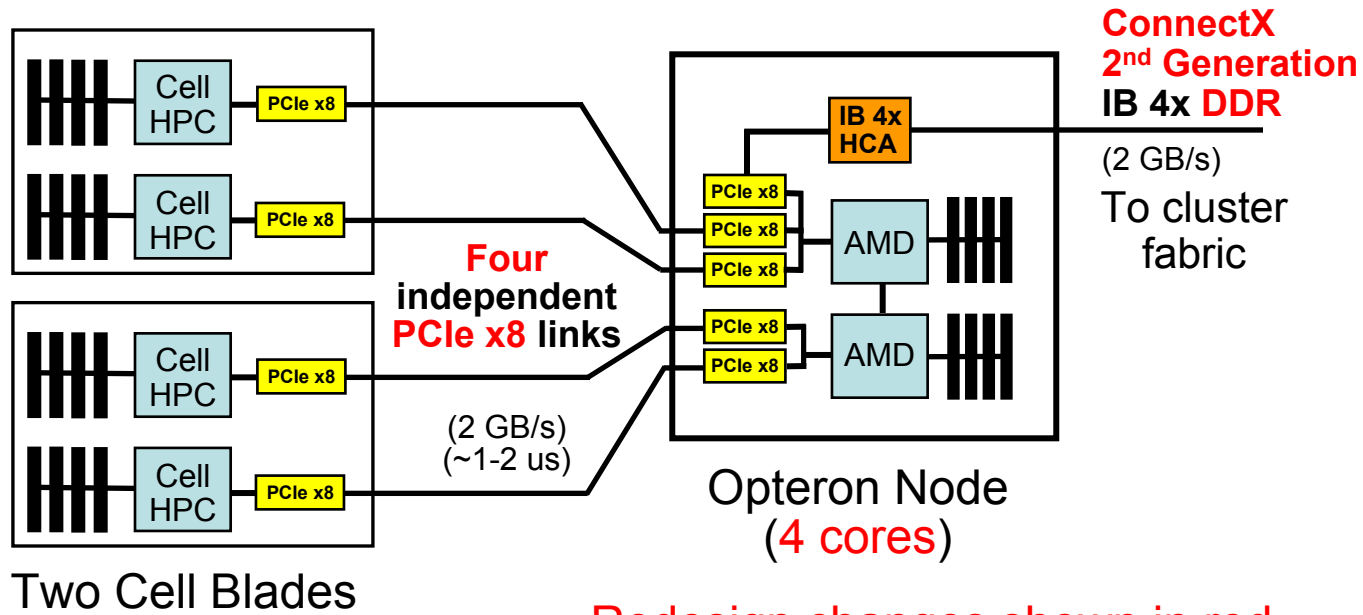


# Roadrunner uses Cells to make nodes ~30x faster



## 400+ GFlop/s performance per hybrid node!

### One Cell chip per Opteron core



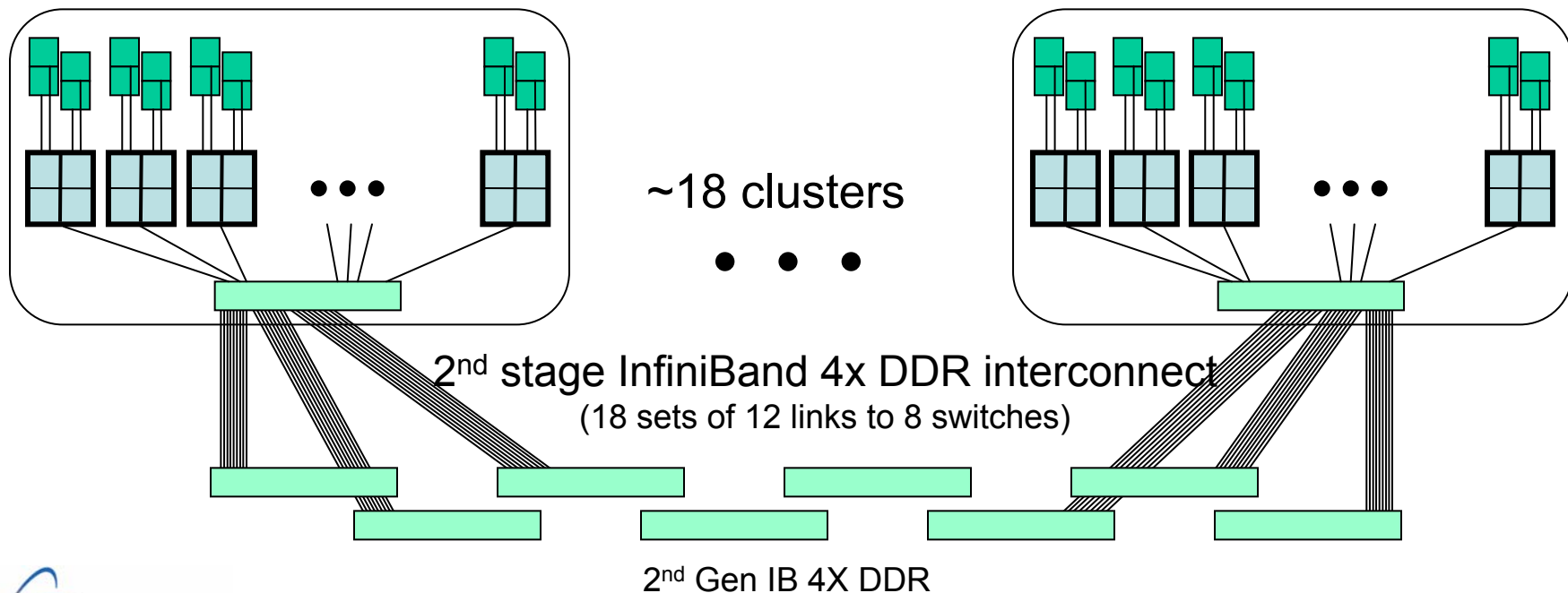
Redesign changes shown in red  
- 4x more BW/flop and better latencies



# Roadrunner is a Petascale system in 2008

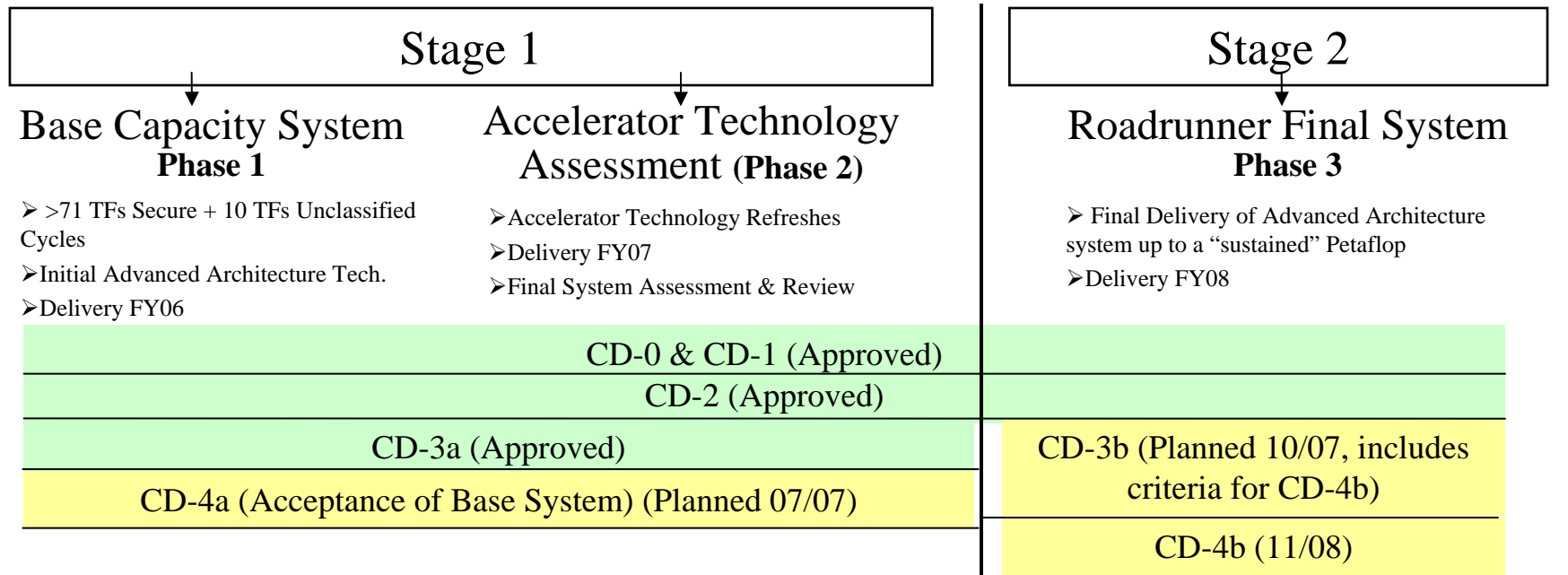
“Connected Unit” cluster  
 192 Opterons nodes  
 (180 w/ 2 dual-Cell blades  
 connected w/ 4 PCIe x8 links)

- ~7,000 dual-core Opterons
  - ~50 TeraFlop/s (from Opterons)
- ~13,000 Cell HPC chips
  - **1.4 PetaFlop/s** (from Cell)





# The Roadrunner procurement is tracked like a construction project via DOE Order 413 w/ NNSA



## Project Status/Milestones

- CD-1 Approval May 4, 2006
- RFP Released May 9, 2006
- Proposals Received June 5, 2006
- Selection made July
- Contract signed Sept. 8, 2006

- Software Architecture Design
- Certification work on Roadrunner

- Option to be executed in early FY08
- Final System Delivery
- Final System Acceptance Test



Operated by Los Alamos National Security, LLC for NNSA



## Roadrunner Phase 3 is an option with planned assessment reviews

---

- Two assessments are planned for October 2007
  - LANL-chartered Assessment committee
  - NNSA ASC chartered independent assessments by HPC experts
- Assessment metrics
  - Performance
    - **Future workload** (e.g. Science@Scale & Advanced Architecture)
    - **Expected Linpack  $\geq 1.0$  PF sustained**
  - Usability and manageability
    - **System management and integration at scale**
    - **API for programming hybrid system**
  - Technology
    - **Delivery of advanced technology**

UNCLASSIFIED

# Programming Roadrunner is tractable



Operated by Los Alamos National Security, LLC for NNSA

**WEAPONS SCIENCE & ENGINEERING  
CAPABILITY REVIEW**



UNCLASSIFIED

## Three levels of parallelism are exposed along with remote offload of an algorithm “solver”

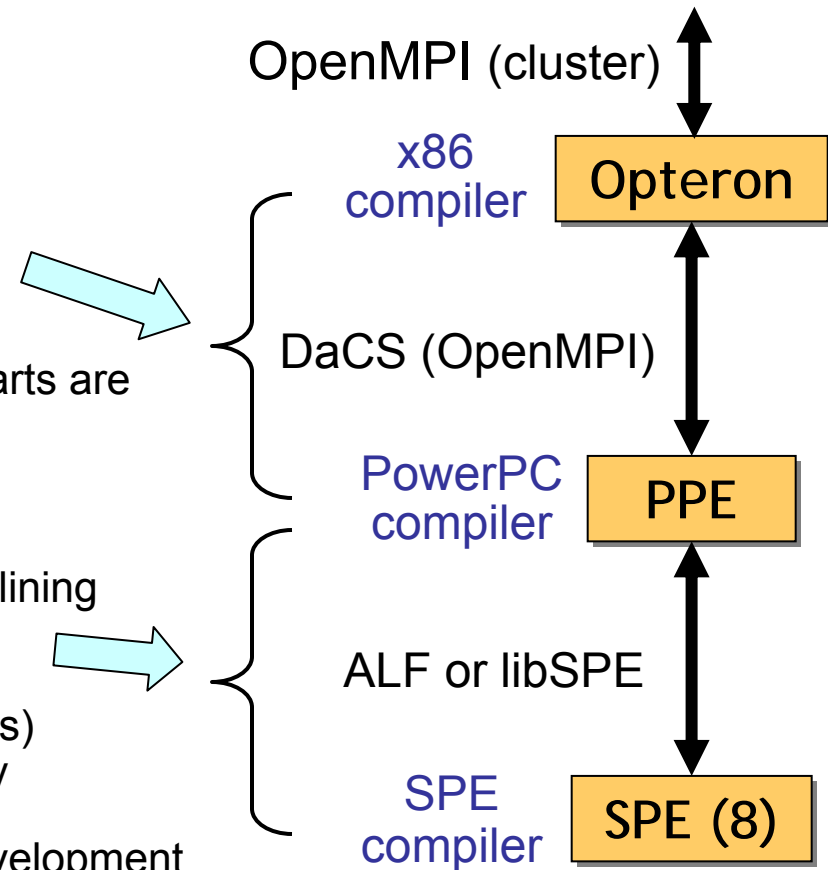
---

- MPI message passing still used between nodes and within-node (2 levels)
  - Global Arrays, IPC, UPC, Global Address Space (GAS) languages, etc. also remain possible choices
  - Additional parallelism can be introduced within the node (“divide & conquer”)
    - Roadrunner does not require this due to it’s modest scale
- Offload large-grain computationally intense algorithms for Cell acceleration within a node process
  - This is equivalent to function offload and similar to client-server & RPCs
    - One Cell per one Opteron core (1:1 process ratio)
    - Opteron would typically block, but could do concurrent work
  - Embedded MPI communications are possible via “relay” approach
- Threaded fine-grained parallelism within the Cell itself (1 level)
  - Create many-way parallel pipelined work units for SPMD on the SPEs
    - MPMD, RPCs, streaming, etc. are also possible
  - Consistent with heterogeneous chips future trends

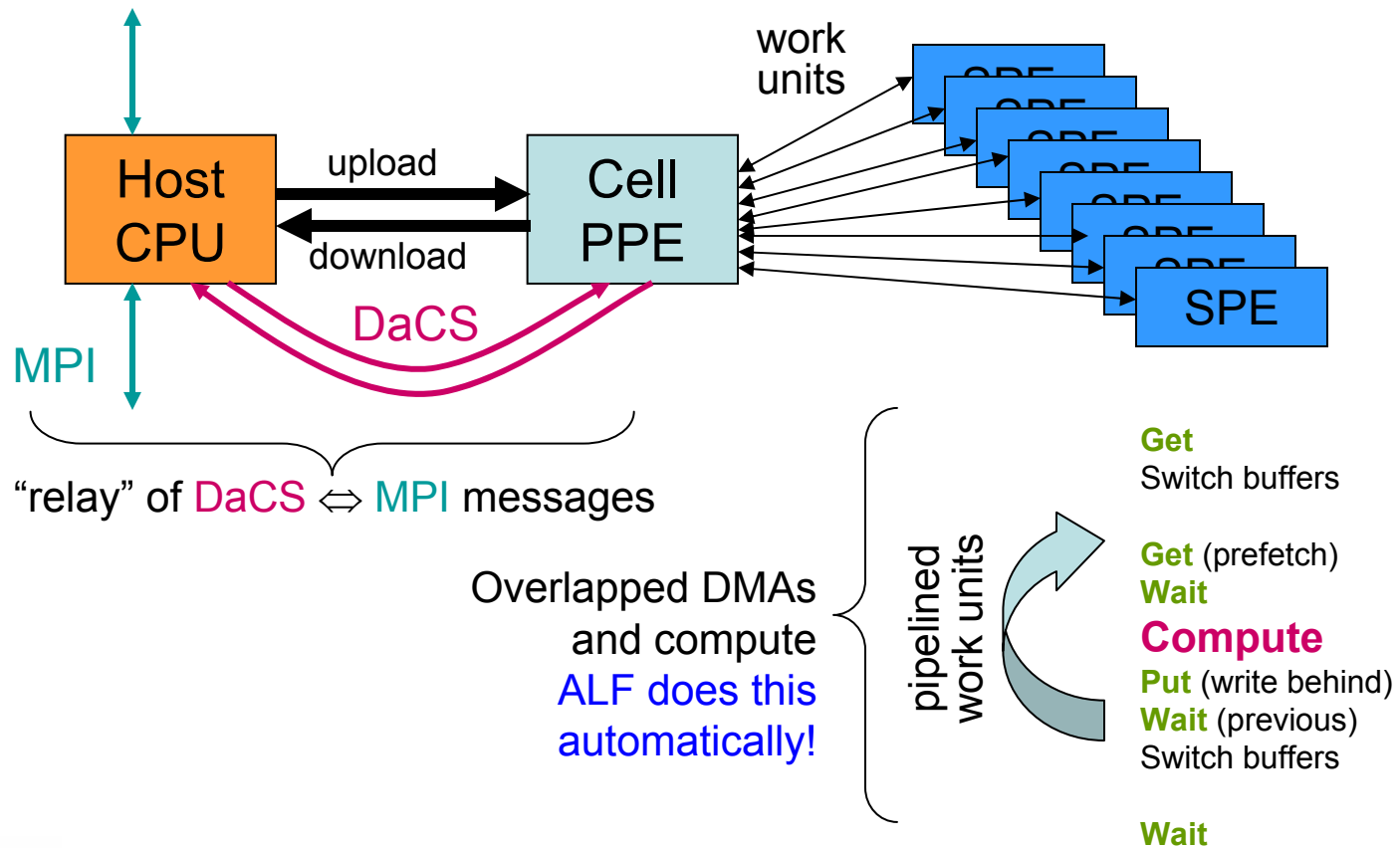
- Considerable flexibility and opportunities exist

## Three types of processors work together

- Remote Communication to/from Cell
  - Data communication & synchronization
  - Process management & synchronization
  - Topology description
  - Error handling
  - IBM developing [DaCS](#) library
  - OpenMPI has proven useful for early “remote computation” prototyping and parts are being ported to PCIe
- Parallel computing on Cell
  - Task data partitioning & work queue pipelining
  - Process management
  - Error handling
  - IBM provides Cell libspe (threads & DMAs)
  - IBM developing ([ALF](#)) Accelerator Library Framework
  - Many 3<sup>rd</sup> party runtimes/languages in development

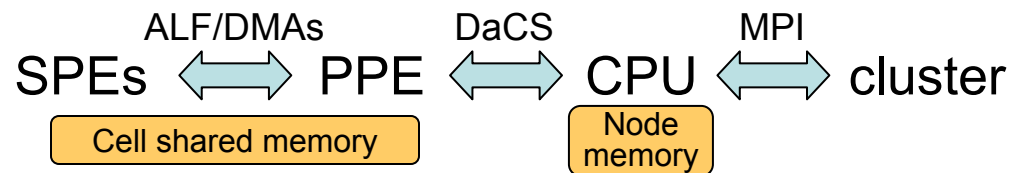


# Three types of processors work together



# Message-passing MPI programs can evolve

- Key concepts:
  - Pair one Cell core with one Opteron core



- Move an entire compute-intensive function/algorithm & associated data onto Cell
  - Can be implemented one function or algorithm at a time
  - Use “message relay” to/from Cells to cluster-wide MPI on host CPUs
- Identify & expose many-way fine-grain parallelism for SPEs
  - Add SPE specific optimizations, many of which are also good on multi-core commodity processors, some are more SPE specific
  - Current SPE programming is limited in C/C++
- Retain main code control, I/O and MPI communications on host CPUs

UNCLASSIFIED

# And what about applications



Operated by Los Alamos National Security, LLC for NNSA

**WEAPONS SCIENCE & ENGINEERING  
CAPABILITY REVIEW**

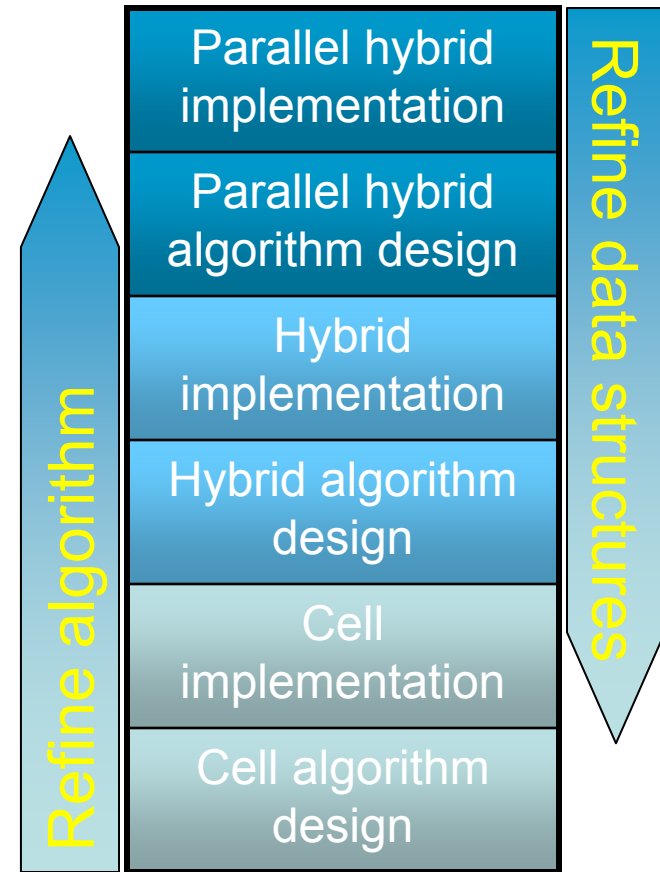
UNCLASSIFIED





## A few key algorithms are being targeted

- Transport
  - PARTISN (neutron transport via Sn)
    - Sweep3D
    - Sparse solver (PCG)
  - MILAGRO (IMC)
- Particle methods
  - Molecular dynamics (SPaSM)
    - Data parallel CM-5 implementation
  - Particle-in-cell (VPIC)
- Eulerian hydro
  - Direct Numerical Simulation
- Linear algebra
  - LINPACK
  - Preconditioned Conjugate Gradient (PCG)



# We are making good progress on applications

Target full application	SPaSM	VPIC	PARTISN		MILAGRO		RAGE	PARTISN, RAGE, Truchas, etc.
Parallel hybrid implementation	8/07 SPaSM in progress	Coding completed	Late due to re-design	Cell cluster version exists	Near coding completion	Near coding completion		
Parallel hybrid design	↑	↑		↑	↑	↑		
Serial Hybrid implementation			On hold	In progress				Hybrid coding completed
Serial Hybrid design							curtailed	Optimizations possible
Cell implementation			Re-design in progress			SIMD coding	↑	
Cell design	CellIMD	VPIC	Re-design completed Sweep3D JK-iagonals	PAL-Sweep3D Domain decomp	Milagro	Milagro rewrite	Research code	CG and GMRES
	Molecular Dynamics (MD)	Particle-in-Cell (PIC)	SPE Threads Deterministic Neutron Transport (Discrete Ordinates, $S_N$ )	SPE Sweeps	re-implement Stochastic Radiation Transport (Implicit Monte Carlo, IMC)	re-design	Eulerian Hydro	Sparse Linear Algebra (Krylov methods)

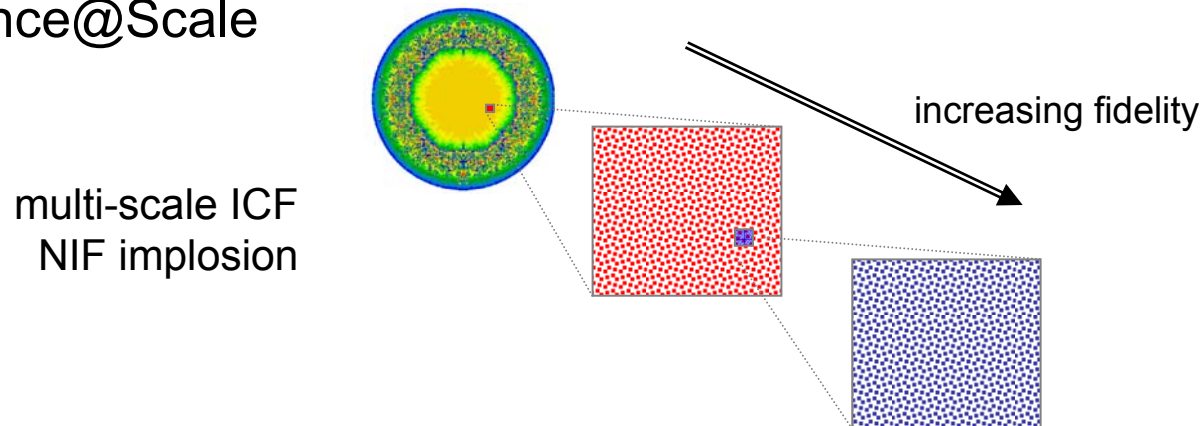
## Roadrunner hybrid implementations are faster

---

- Speed ups so far range from 1x (disappointing: redesign & more optimizations) to 9x (very good)
  - Reference performance is taken as the performance on a 2.2GHz Opteron core or cluster of such Opterons
  - Science codes are fairing the best
- Extensive performance modeling with key measurements on hardware prototypes will be used to project final Roadrunner performance of these applications
  - AAIS machine has IB-connected Cell blades
  - A few Cells are connected to Opterons via PCIe at IBM Rochester site
  - IBM is building a ConnectX IB cluster for testing at Poughkeepsie
  - A few prototype hybrid nodes with current Cell chips will be available in late August in Rochester
  - A couple of new Cell HPC chips are available at IBM Austin in test rigs
- Much work is yet to be accomplished by the October Assessments

## Roadrunner targets two application areas

- Science@Scale
  - Targeting VPIC & SPaSM codes for weapons science
  - Cross-validate physics models at overlapping resolutions
  - Run at Petascale
    - ~75% of machine, for 1½ to 2 week durations, is minimally needed for each VPIC ICF study
  - Guy Dimonte will talk next about VPIC and its role for Science@Scale



## Roadrunner targets two application areas

---

- Advanced Architecture for algorithms and applications
  - Targeting unclassified transport algorithms initially
  - Provide faster solutions or improved accuracy
  - Advanced parallel hybrid algorithms and application development
  - Demonstrate a path to incrementally updating existing ASC integrated codes
    - Target key physics or simulation uncertainties, not general DSW/baselining usage
    - Potentially target 3D safety

More information is available on the LANL Roadrunner home page

---

<http://www.lanl.gov/roadrunner/>

Roadrunner Architecture

Other Roadrunner talks

Computing Trends

Related Internet links

UNCLASSIFIED

**WEAPONS** *SCIENCE & ENGINEERING*  
*CAPABILITY REVIEW*

The End