

Roadrunner: Science, Cell and a Petaflop/s

Fall Creek Falls Conference 9/9/08

Andy White

Los Alamos National Laboratory

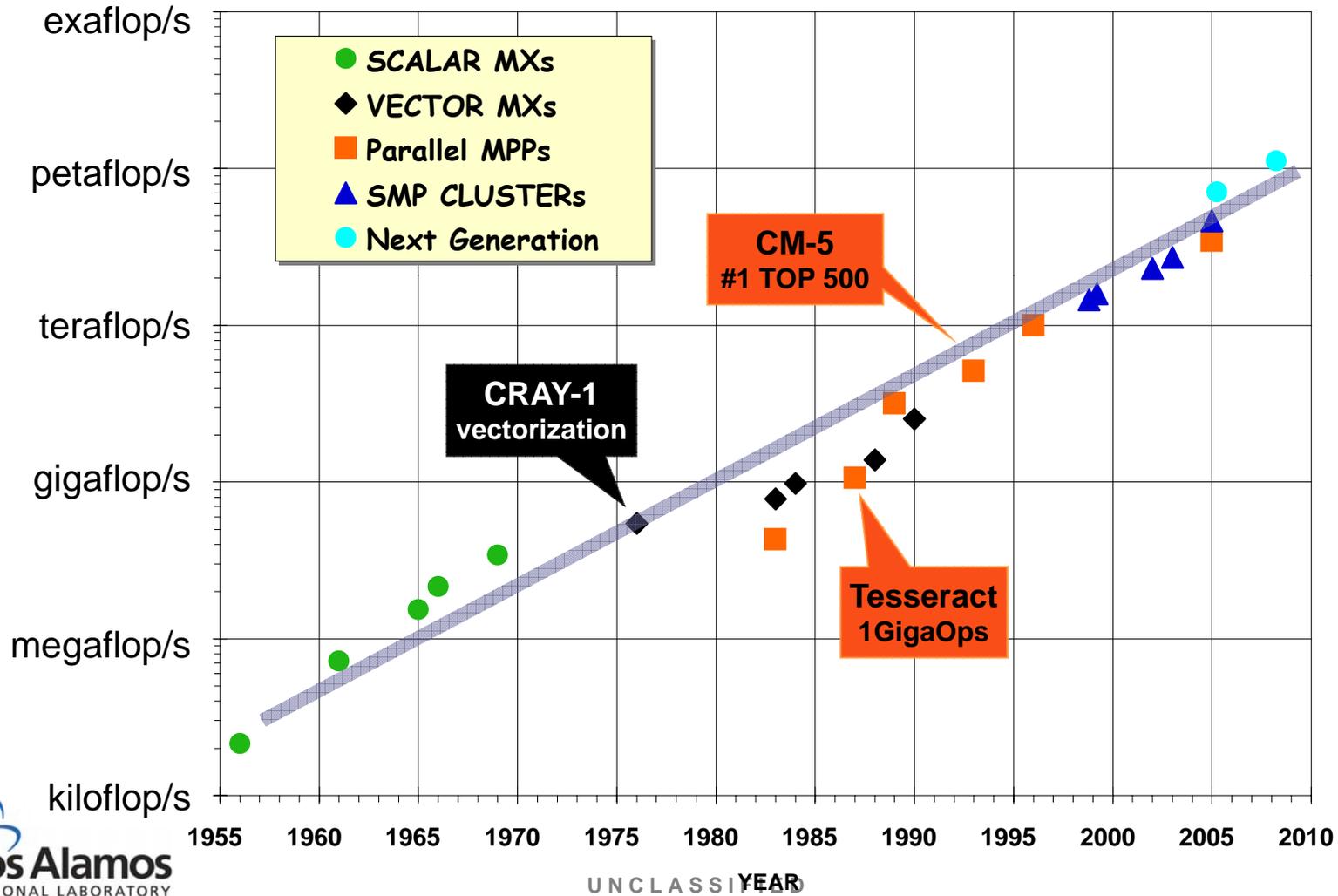


UNCLASSIFIED

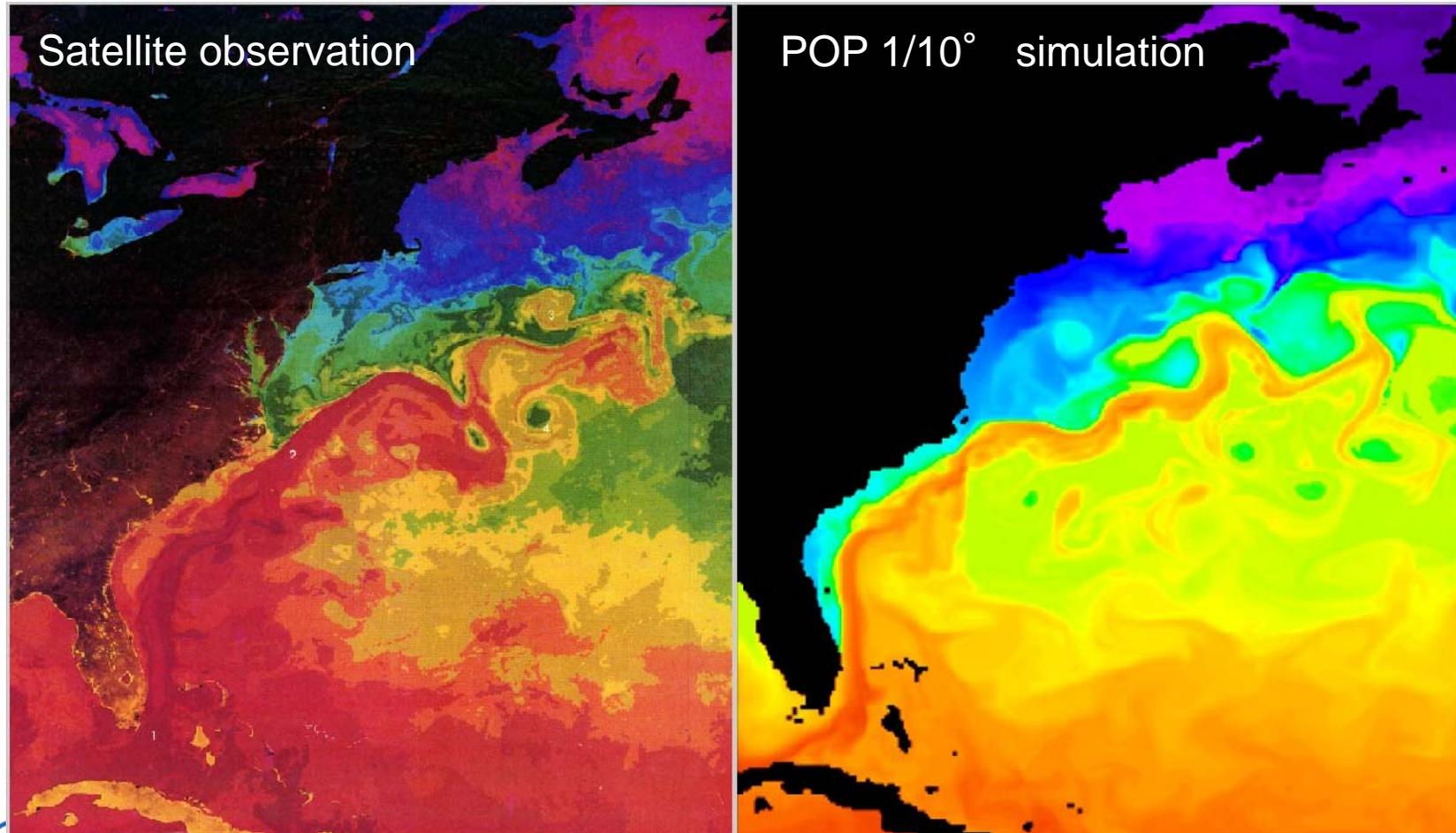
Operated by Los Alamos National Security, LLC for NNSA



Hybrid, many core chips are changing our world, again.

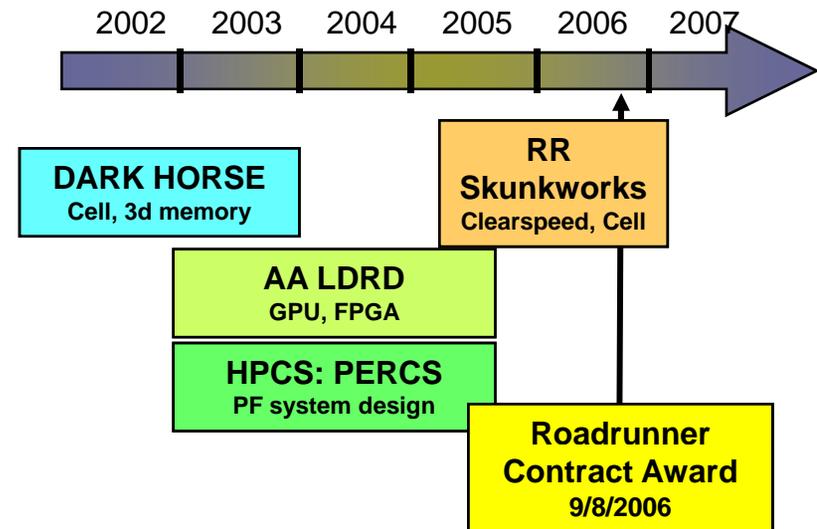


Transformational technologies are a catalyst for opportunity.



After almost six years, an accelerator-based petascale supercomputer is becoming a reality.

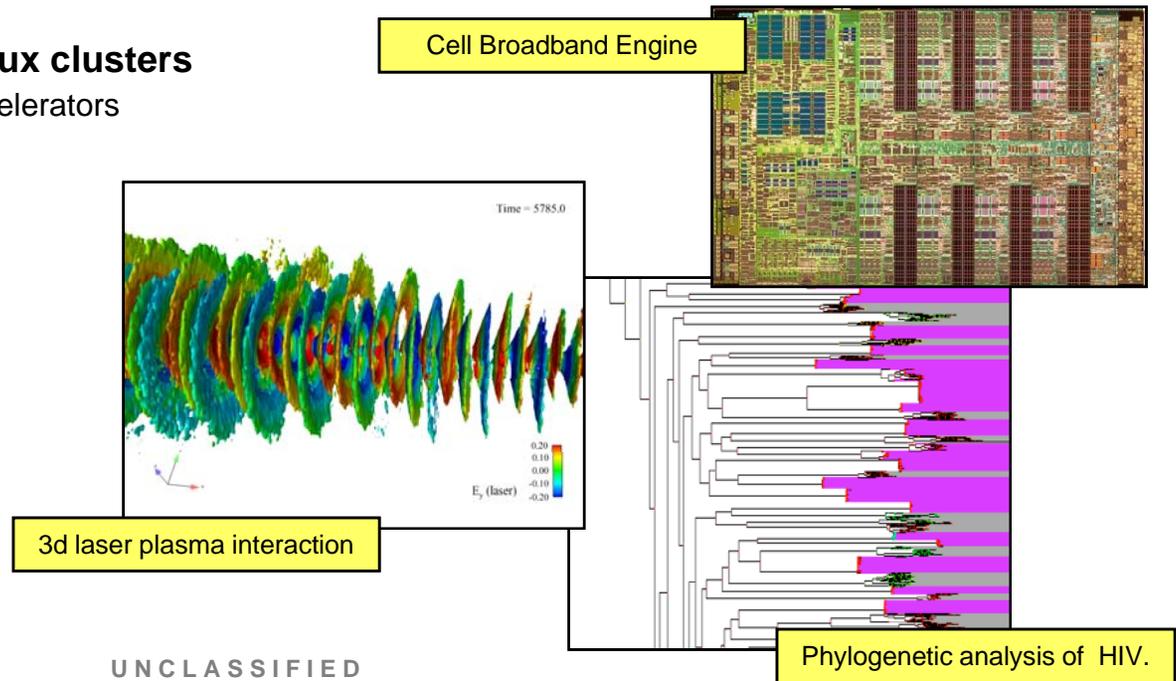
- **Phase 1: Provide a large, stable capacity resource for the weapons program**
 - ✓ The base system is contributing to classified operations now
 - ✓ The system more than doubles our capacity computing for the weapons program: 71 Teraflop/s
- **Phase 2: Examine the long-term viability of Roadrunner for weapons program**
 - ✓ Prototype hardware and early software drops
 - ✓ Final System Technical Assessment (October 17-19, 2007)
 - ✓ CD3b signed (12/2007)
- **Phase 3: Provide a petascale resource for the weapons program**
 - ✓ Contract Option to manage risk
 - Science at scale (e.g. plasma physics)
 - Advanced architecture for integrated codes
 - ✓ Sustained Petaflop/s Linpack Final System



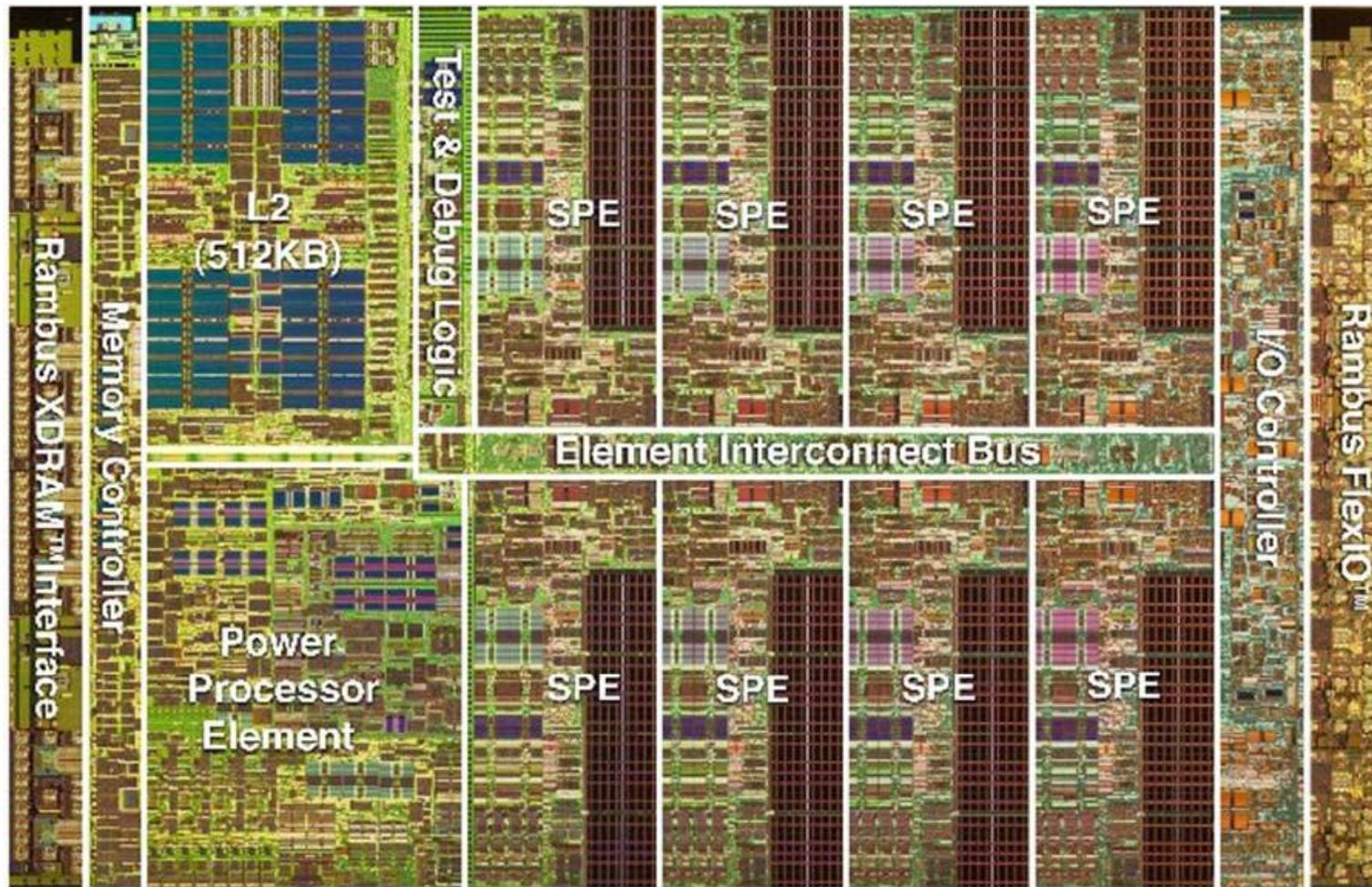
Rockstar games

Roadrunner at a glance

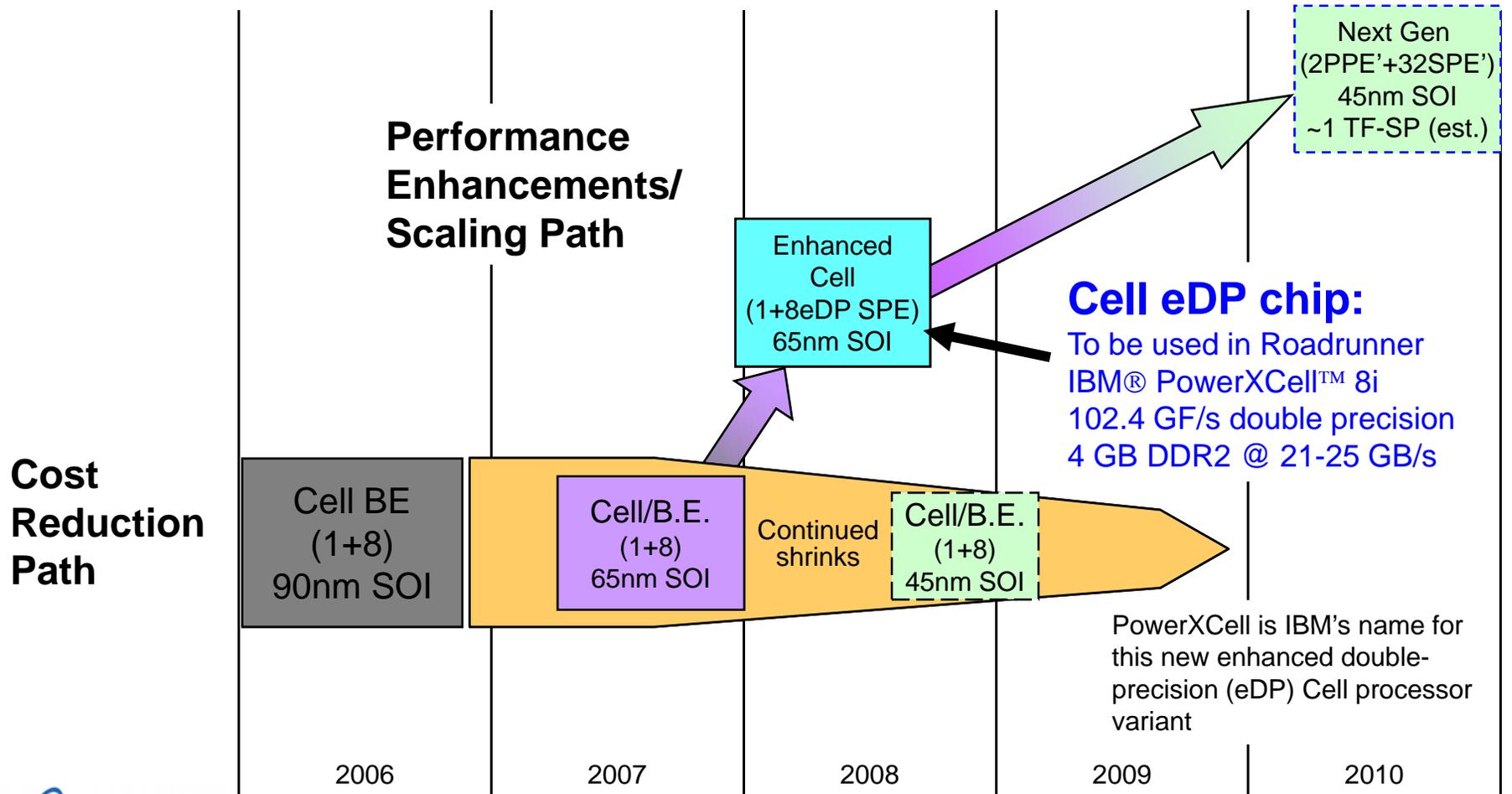
- **PowerXCell 8i provides floating point performance**
 - Derivative of Cell chip in PlayStation 3
 - 108.8 Gigaflop/s peak
- **Triblade nodes are interconnected by InfiniBand 4x DDR**
 - 4 PowerXCell 8i
 - 2 AMD dual core Opteron
- **System is a cluster of 17 Linux clusters**
 - 12,240 IBM PowerXCell 8i accelerators
 - 3,060 nodes (Triblades)
 - 1.376 Petaflop/s peak
- **98 TB aggregate memory**
 - 49 TB Opteron
 - 49 TB Cell
- **Other facts:**
 - 5200 ft²
 - 500,000 lbs.
 - 55 miles of IB cables
- **First-of-a-kind hybrid, many core supercomputer**
- **First system to achieve a sustained petaflop/s (10^{15}): 1.026 Petaflop/s**
- **Most power efficient *supercomputer*: 437 Megaflop/s per watt**



IBM has developed the (hybrid) PowerXCell 8i with enhanced double precision and memory access.



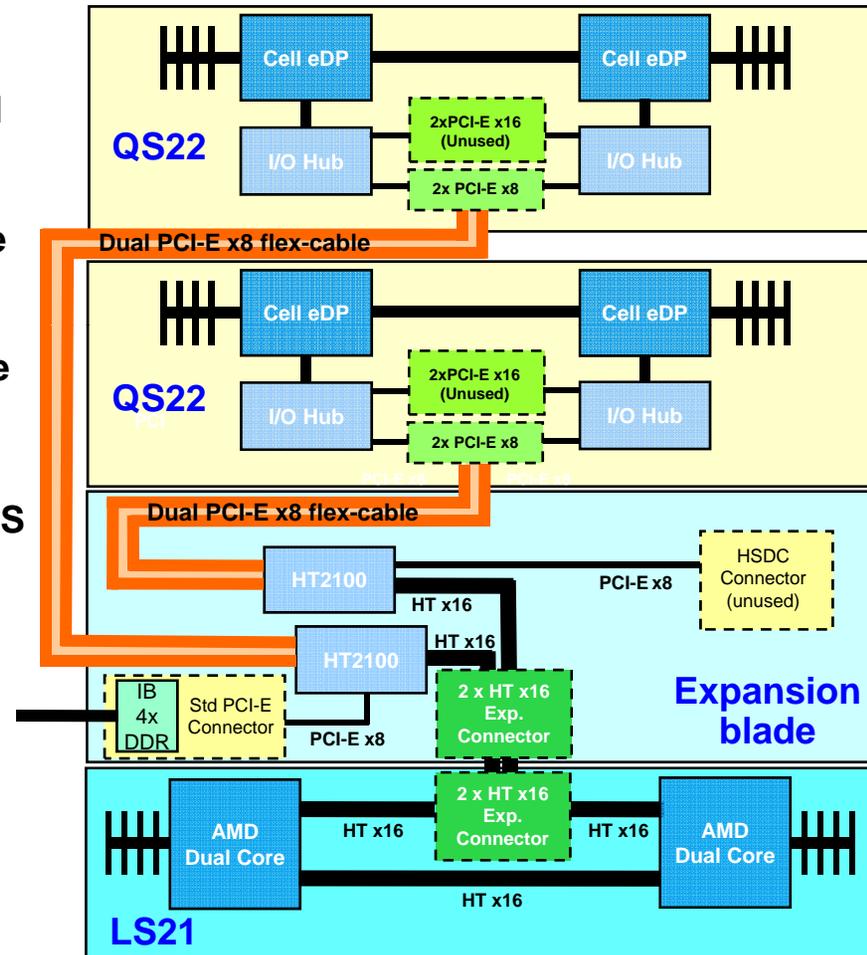
Cell Broadband Engine™ Architecture (CBEA) Technology Competitive Roadmap



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 Optron blades

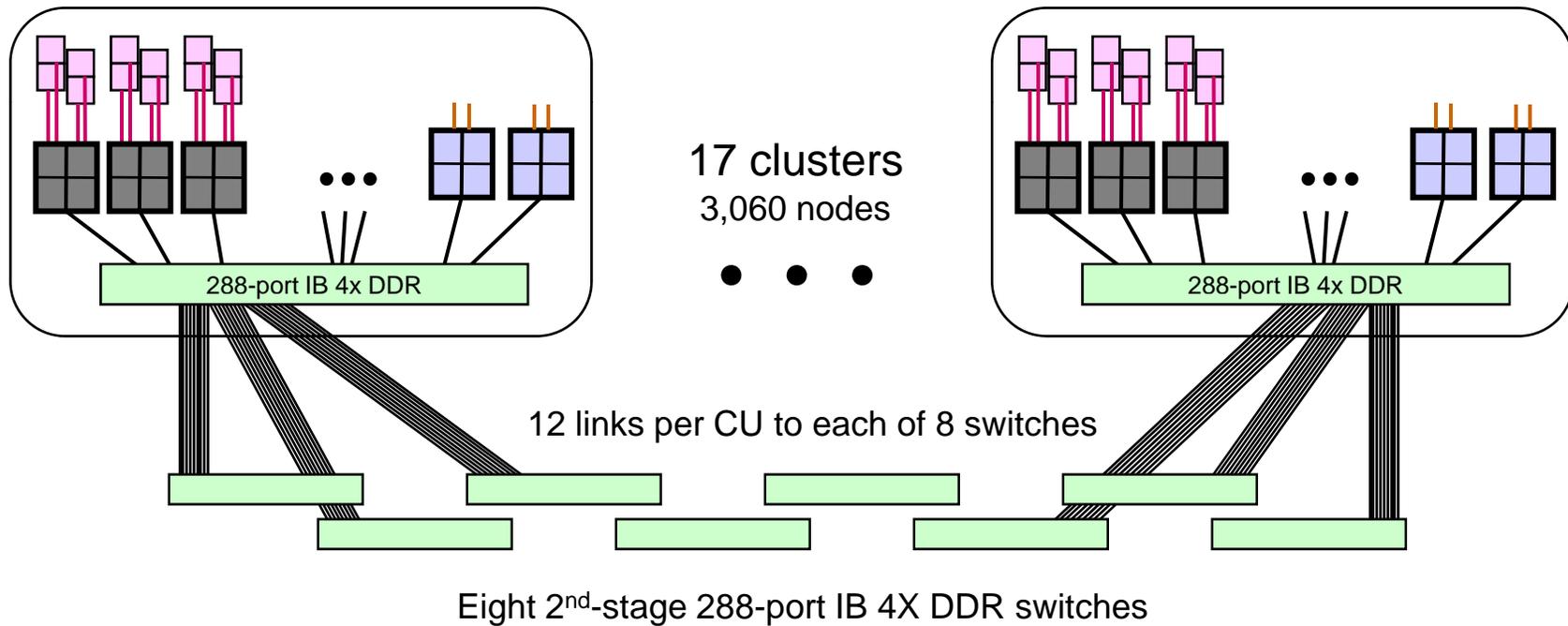
- **QS22** is the newly announced IBM Cell blade containing two new enhanced double-precision (eDP/PowerXCell™) Cell chips
- Expansion blade connects two **QS22** via **four PCI-e x8** links to **LS21** & provides the node's ConnectX IB 4X DDR cluster attachment
- **LS21** is an IBM dual-socket Optron blade
- 4-wide IBM BladeCenter packaging
- Roadrunner Triblades are completely diskless and run from RAM disks with NFS & Panasas only to the LS21



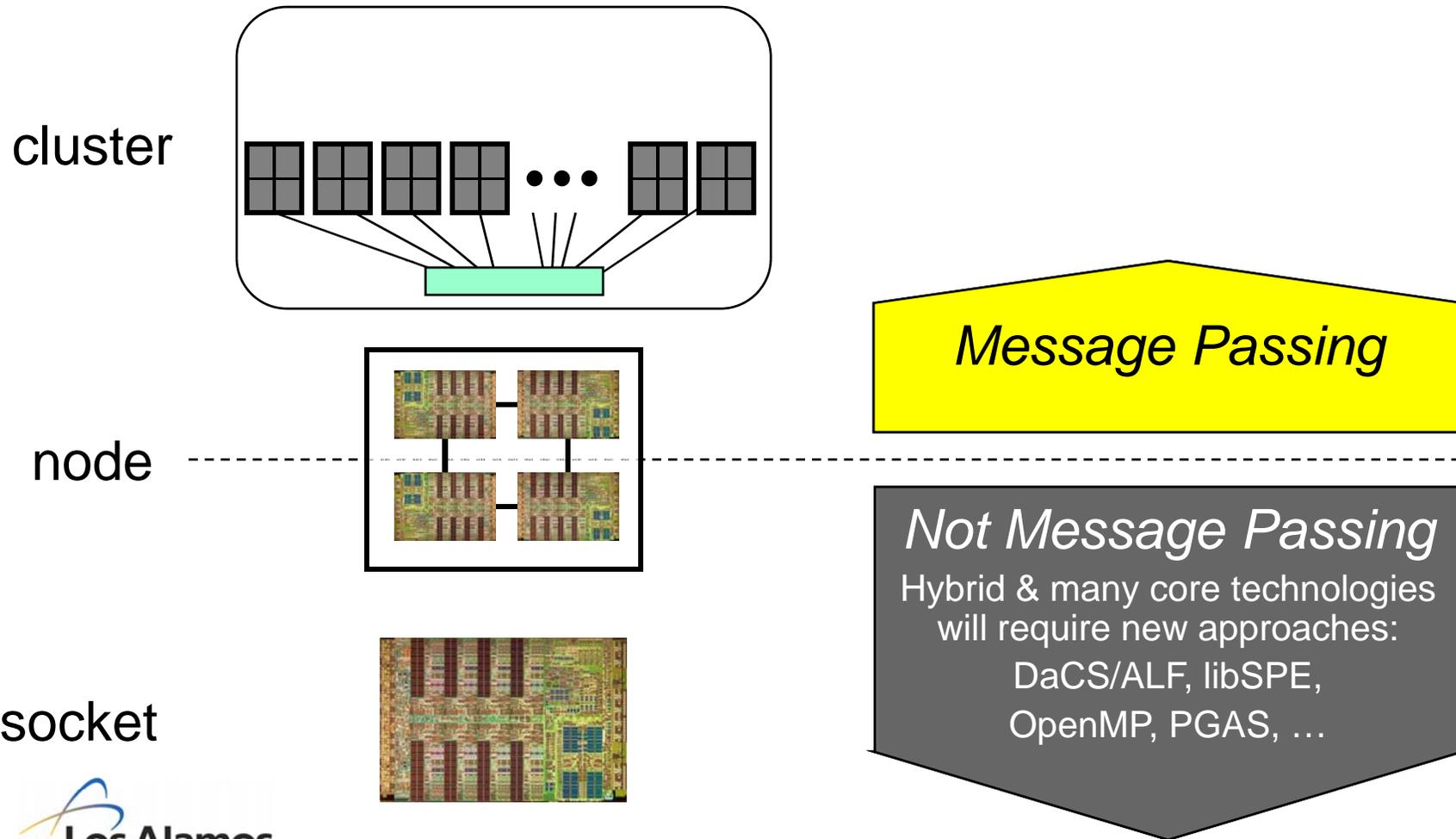
Roadrunner is a hybrid² petascale system of modest size.

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

12,240 Cell eDP chips \Rightarrow 1.3 PF, 49 TB
6,120 dual-core Opterons \Rightarrow 50 TF, 49 TB



Roadrunner is not MPI everywhere.

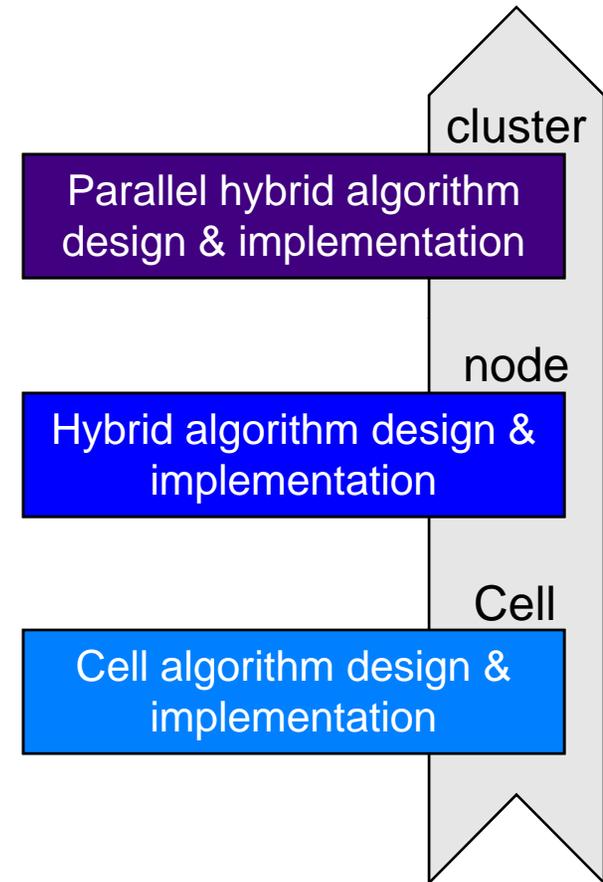


Message Passing

Not Message Passing
Hybrid & many core technologies
will require new approaches:
DaCS/ALF, libSPE,
OpenMP, PGAS, ...

We have focused on important application codes.

<i>Code</i>	<i>Description</i>
VPIC <i>(8.5K lines)</i>	Fully-relativistic, charge-conserving, 3D explicit particle-in-cell code.
SPaSM <i>(34K lines)</i>	Scalable Parallel Short-range Molecular Dynamics code, orig. developed for the CM-5.
Milagro <i>(110K lines)</i>	Parallel, multi-dimensional, object-oriented code for thermal x-ray transport via Implicit Monte Carlo on a variety of meshes.
Sweep3D <i>(2.5K lines)</i>	Simplified 1-group 3D Cartesian discrete ordinates (Sn) kernel representative of the PARTISN neutron transport code.



A most difficult test was proving that real applications would perform well on Roadrunner: October 17-19, 2007.

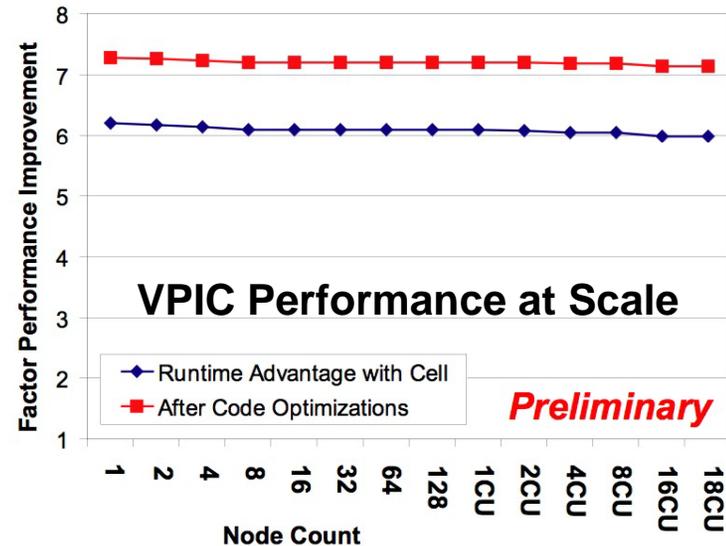
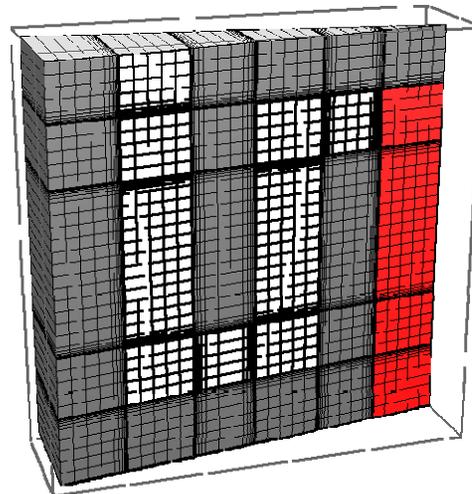
Advanced Algorithms and Applications team

Application	Class	eDP Cell
<i>SPaSM</i>	full app	4.5x
<i>VPIC</i>	full app	9x
<i>Milagro</i>	full app	6.5x
<i>Sweep3D</i>	kernel	9x

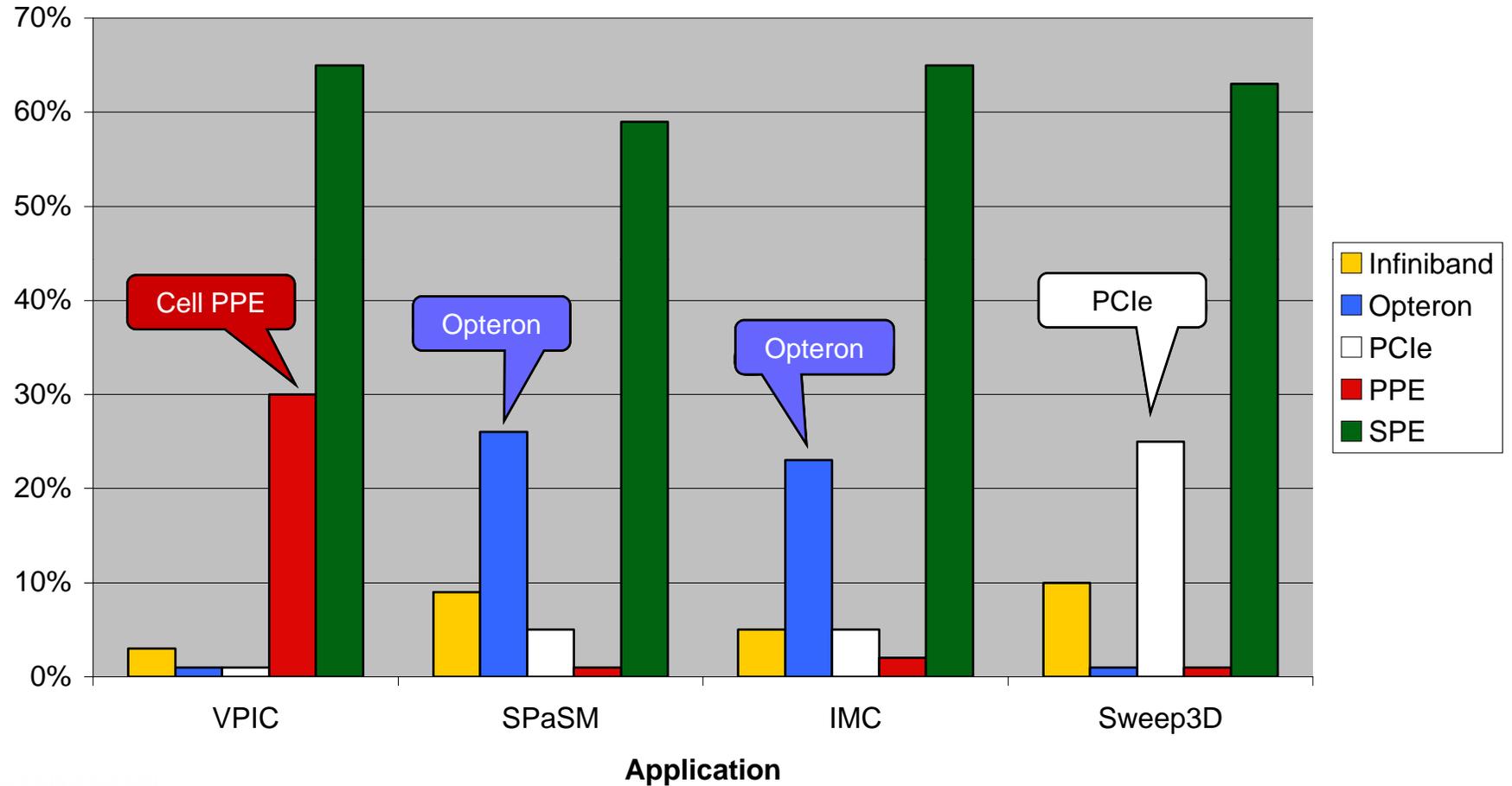
Performance and Architecture Laboratory

Double bend test problem for Milagro

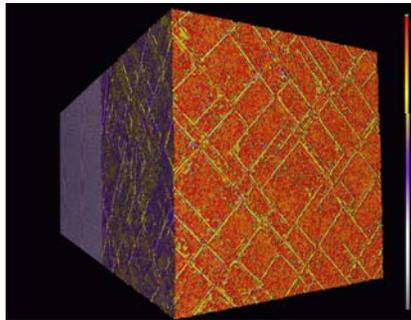
- RZWedge mesh
- optically thick cells (grey)
- optically thin cells (white)
- Hot source (red)



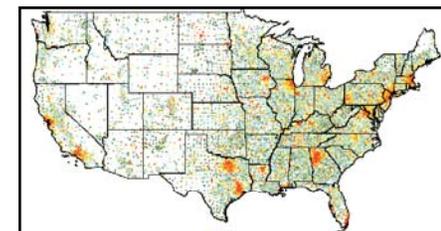
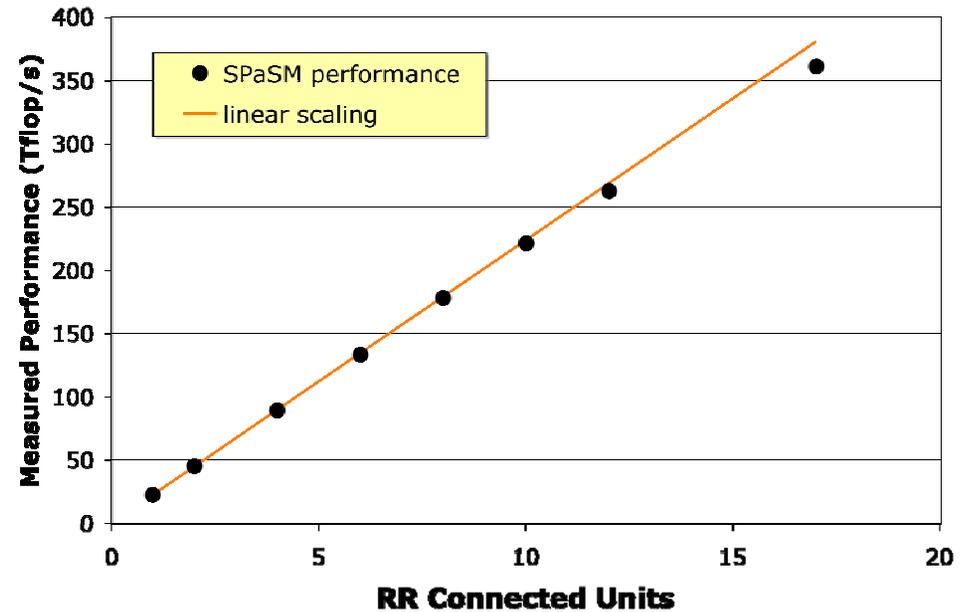
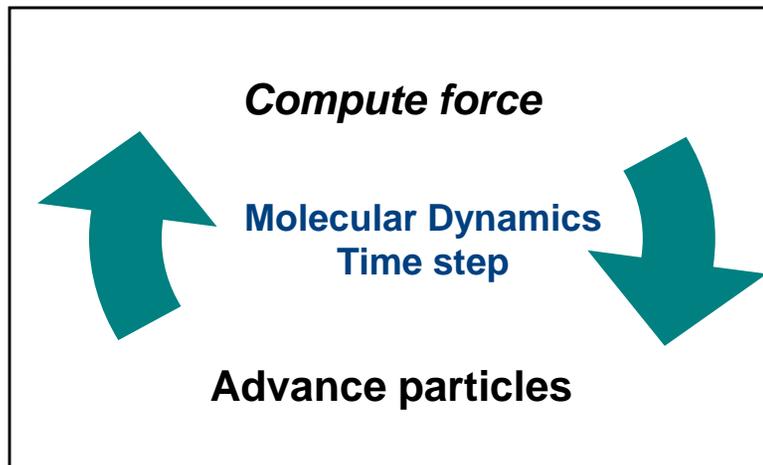
Roadrunner architecture is flexible.



SPaSM is a framework for materials science research as well as epidemiology, fluid instabilities and turbulence.

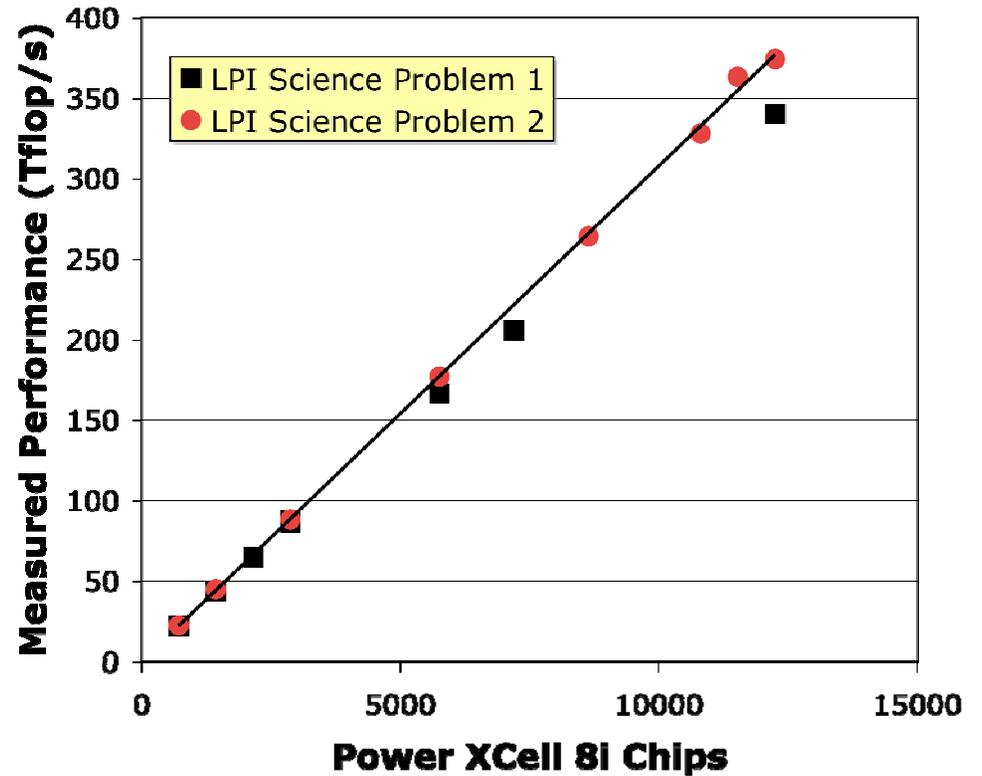
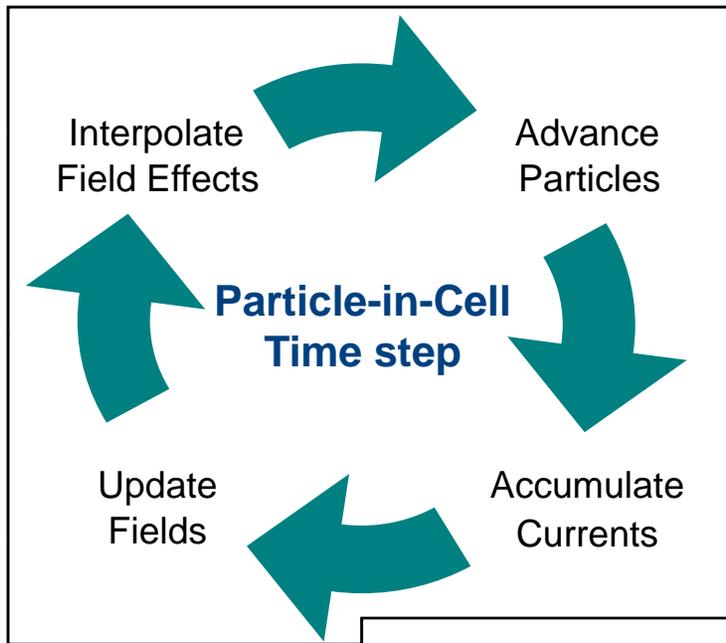


Shock compression of metals

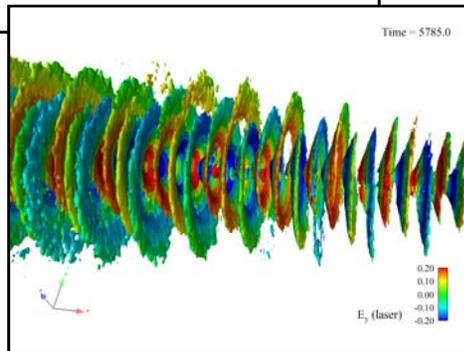


Pandemic

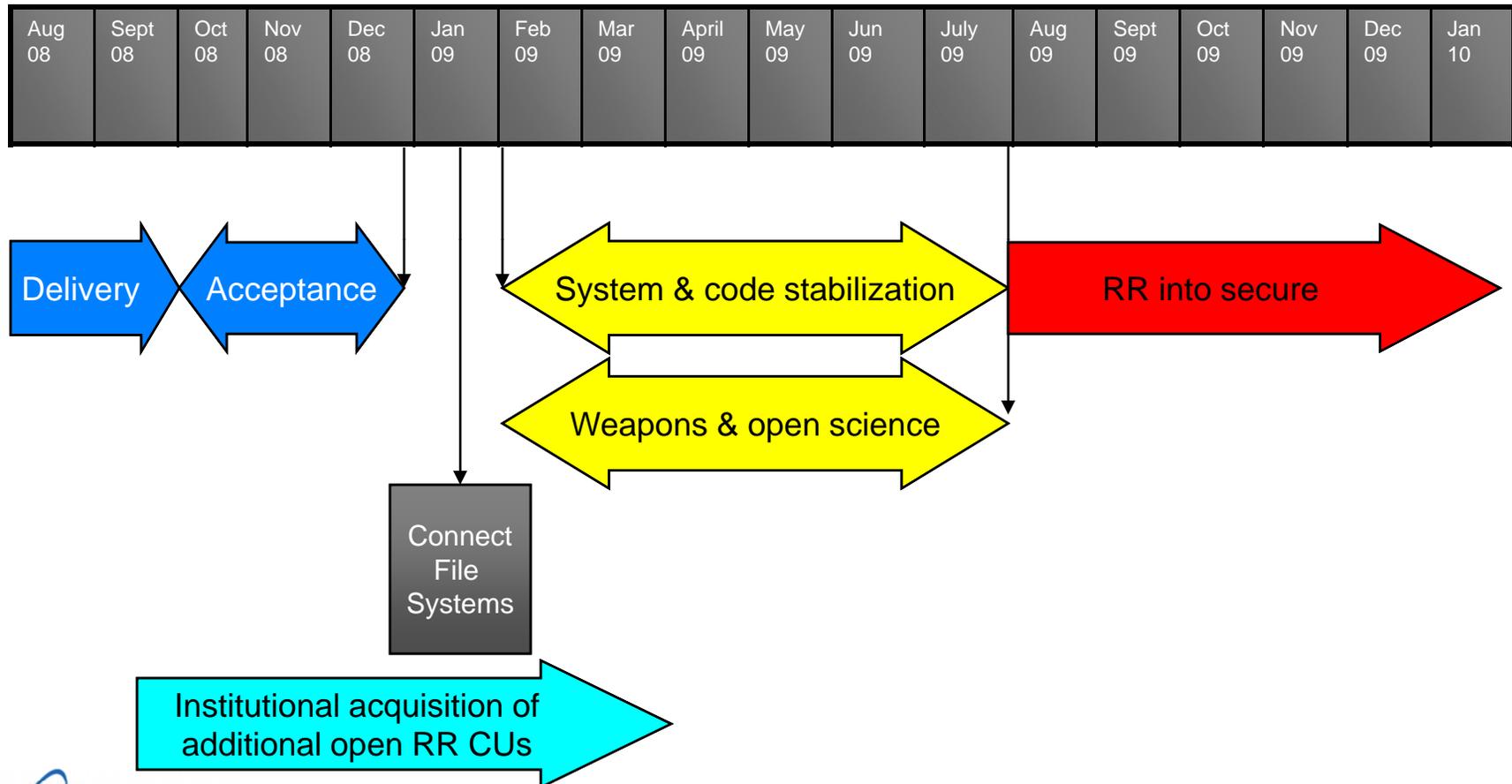
VPIC will address grand challenges in plasma physics.



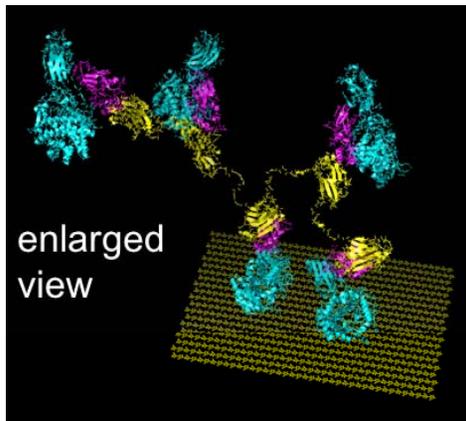
3d laser plasma interaction



We are planning for open science on Roadrunner.

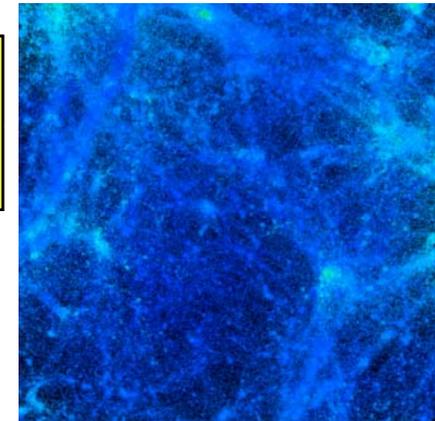


Five RR open science projects will develop new codes.

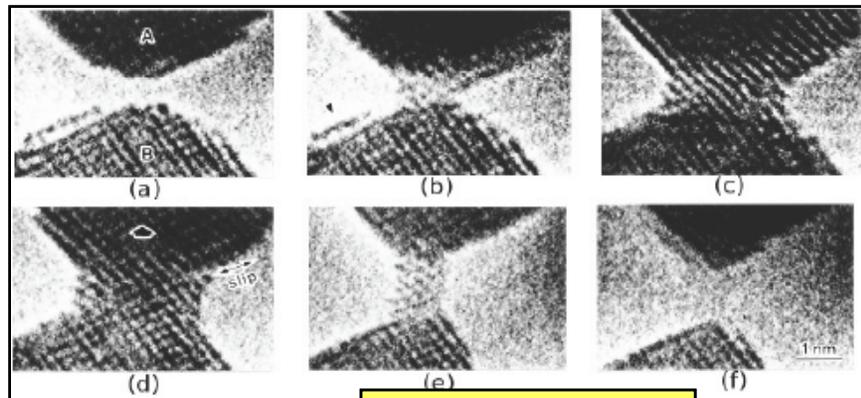
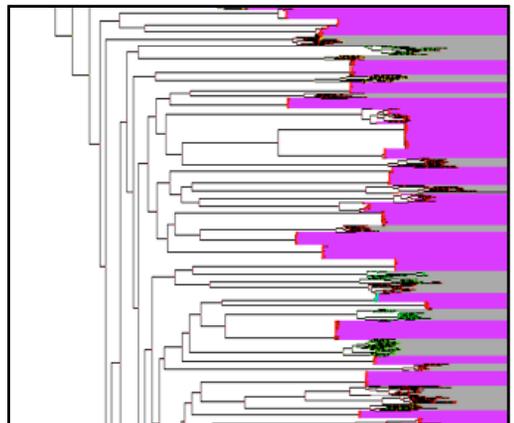


Dynamic molecular simulation of breakdown of cellulose for biofuels.

Cosmological simulation of large scale structure of the universe



Phylogenetic analysis of the evolution of acute HIV infection.



Long-time evolution of the formation & deformation of metallic nanowires

A cycle of change in high performance computing, perhaps, but it's much more interesting than that.

- **1970s: vectorization transforms high performance computing**
- **1990s: large-scale, distributed memory parallelism transforms high performance computing**
- **2010s: hybrid many-core will transform high performance computing**
- **But there are lots of things to think about ...**
 - Now, all programmers have to embrace parallelism
 - *NAG*: teach the world how to program in parallel.
 - *Intel*: fundamental change in the way developers think about programming
 - *IBM*: Software Development Kit for multicore acceleration
 - Errors: hard, soft and silent will become more prevalent
 - 10^{15} flop/sec, 10^7 sec/year, 10^8 transistor/Cell
 - Power estimates for an Exaflop/s system are in 100s of MW/year
 - Life cycle costs dominated by the cost of power
 - Many existing codes were designed to minimize computation, but that may not be an effective design principle in the future
 - Data intensive supercomputing may be a different design point than numerically intensive supercomputing