Sweep3D (Sn transport)
& other key Roadrunner applications

Ken Koch

Roadrunner Technical Manager

Scientific Advisor, CCS-DO

November 29, 2007

What is Deterministic Sn Transport?

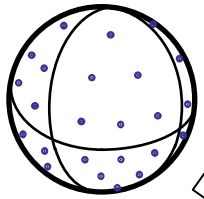
Particle density distribution (e.g. neutrons): $N(\underline{r},E,\underline{\Omega},t)$

position, energy, direction, time

velocity

Discrete directions $\underline{\Omega}_m$

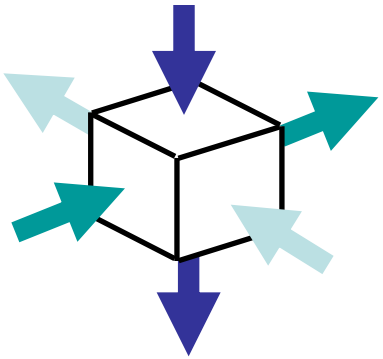
- S_6 has 6 directions per octant
- 48 angles total



Energy treated as histogram bins

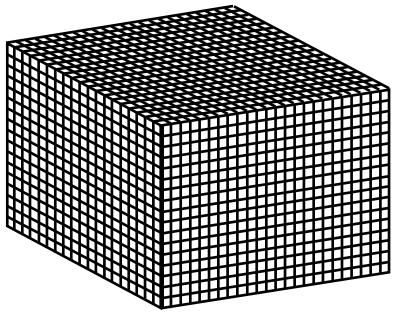
- e.g. 10 -100 bins is representative

Conservation equations on a mesh for each $\underline{\Omega}_m$ direction



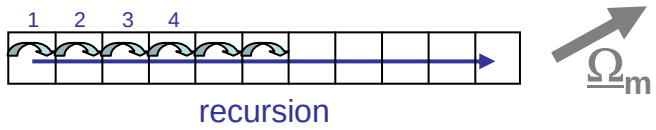
volumetric terms plus
3 inflows & 3 outflows
3D neighbor dependence
(dependant on $\underline{\Omega}_m$ direction)

XYZ mesh
(e.g. 400^3)

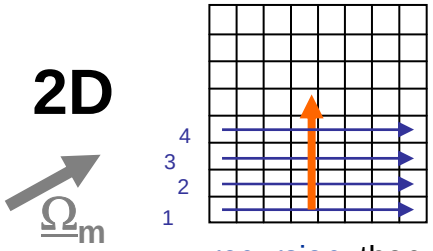


Inflows & outflows cause data dependencies which are handled by **sweeping**

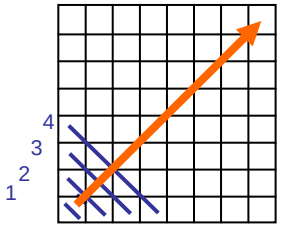
1D



2D



recursion, then recursion



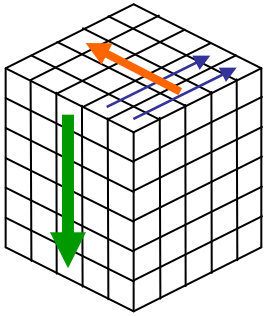
concurrent, then recursion

Cray YMP did this

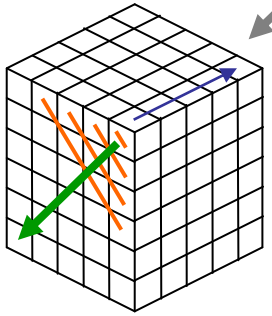
Sweep3D does this

3D

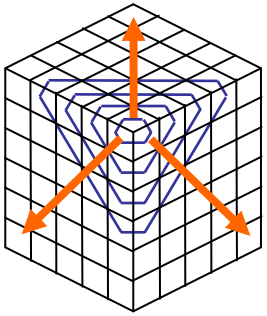
Ωm into corner



recursion, then recursion, then recursion



recursion, then concurrent, then recursion



2D-concurrent, then recursion

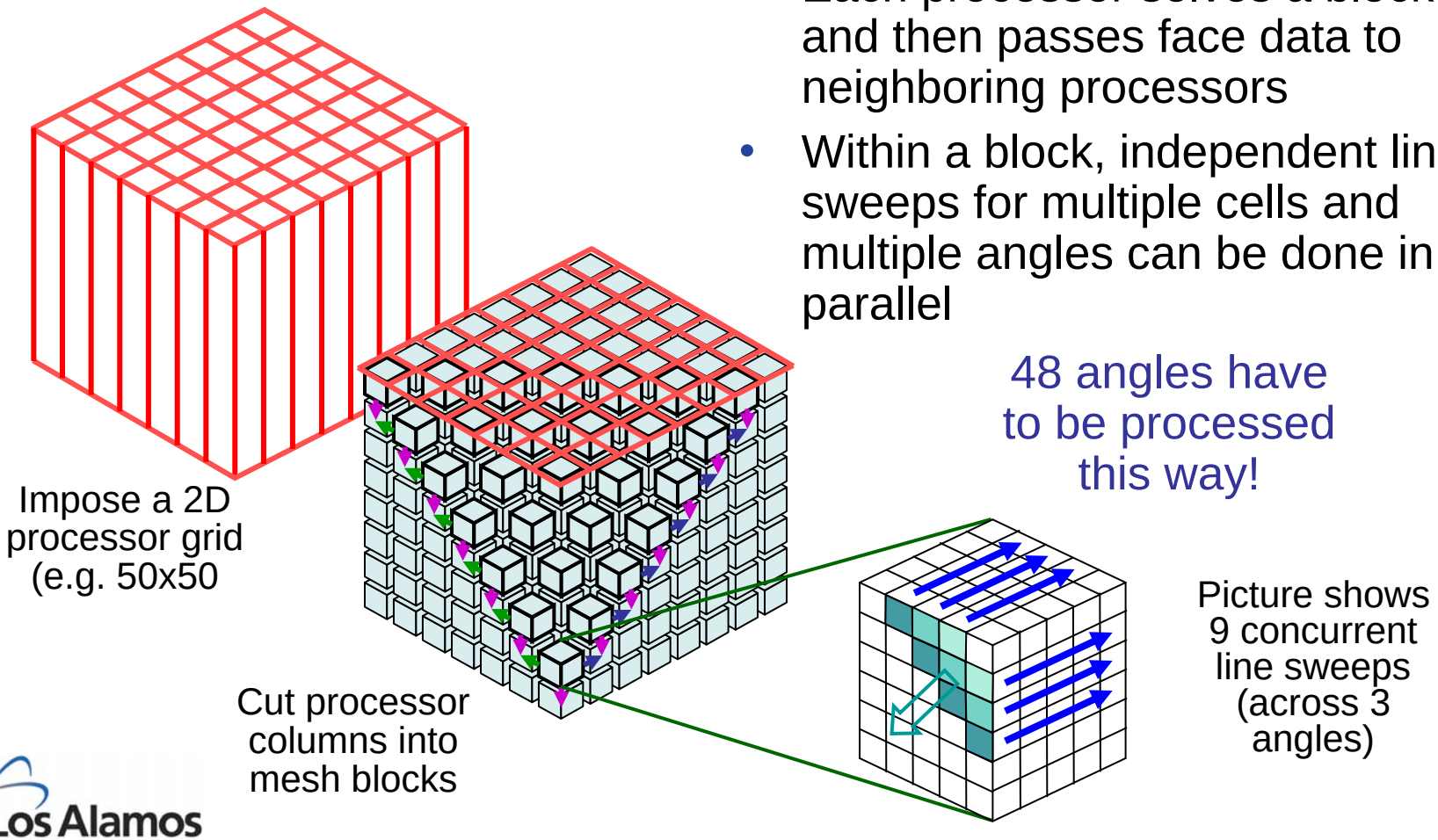
CM-5 did SIMD on a 2D top projection plane

Sweeps are a direct solve technique and avoid iteration

(matrix equivalent is 48 lower-triangular tri-banded blocks with non-banded 48-way coupling-terms treated by iteration)

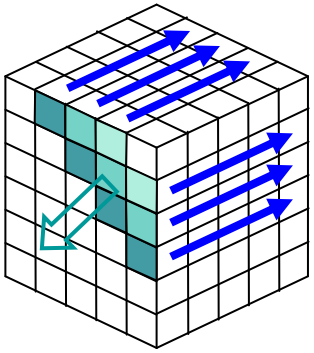
Sweep3D pipelines work across a 2D processor grid using mesh blocks

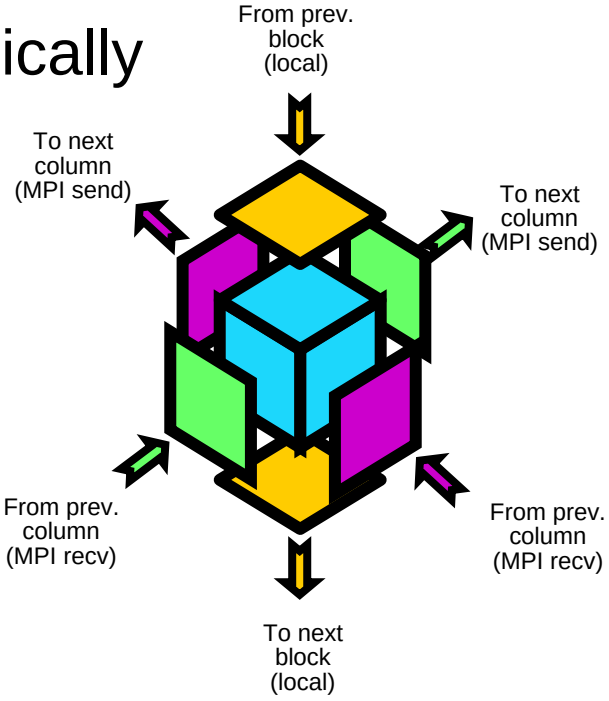
- Each processor solves a block, and then passes face data to neighboring processors
- Within a block, independent line sweeps for multiple cells and multiple angles can be done in parallel



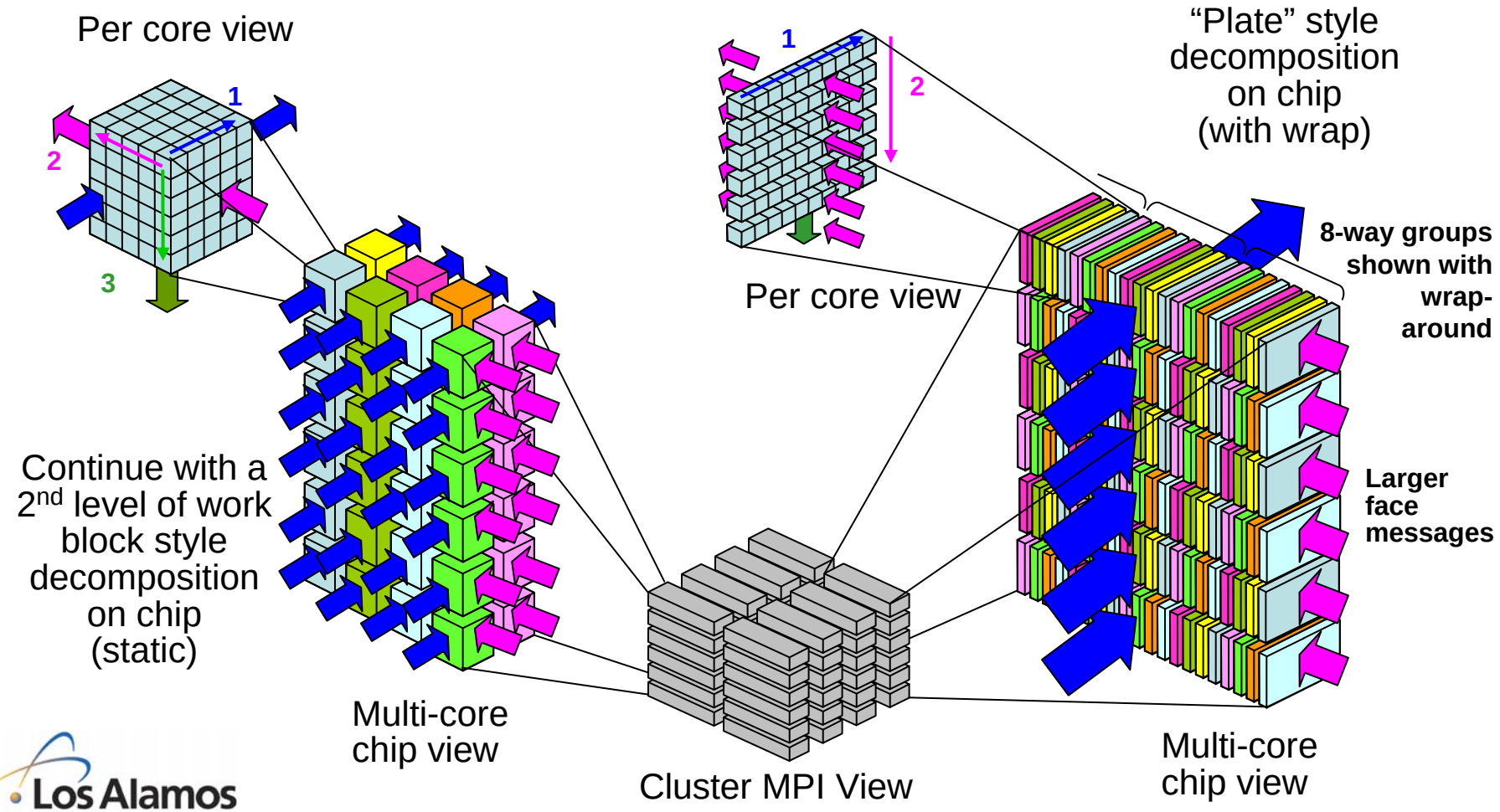
Work blocks are small to maintain pipelined parallelism efficient across processors

- Pipelining exposes parallelism
  - More MPI parallelism requires smaller work blocks and/or larger problems
  - Angles can also be used to lengthen the pipeline
- Work blocks may be 5×5×(2-20) typically
  - Short loops
  - Only 50 to 500 updates per block
  - Only 10 to 100 cells per face





Alternative node-level parallelism approaches are possible and will be needed for future chips (e.g. Cell)



Application characteristics & observations

- Average update time of 30-100 ns
  - 40 to 50 flops per update
  - 10 to 15 DP load/stores per update
  - ~1 GF/s and ~2 GB/s
- Data size is I×J×K×M×G×2 for time-dependant problems
  - 1 TB to 20TB aggregate, just for solution array
  - 200MB to 1GB (per core), plus cross section data
  - Energy & angle domains currently are under-resolved
- MPI transfers
  - 2K to 4K byte messages
  - 4 messages (2 recv + 2 send) per 100 to 200 updates (4 to 10 us)
- Overall performance is a balance of local compute rate and messaging rates
  - Asynchronous send/recv helps overlap compute & communication (when it actually works right)
  - Sweep progresses in a “data-dependant” synchronized manner
  - 100K – 1M way parallelism stresses pipelined parallel efficiency for current sweep approach
- SIMD processing requires word-level gather/scatter to be effective

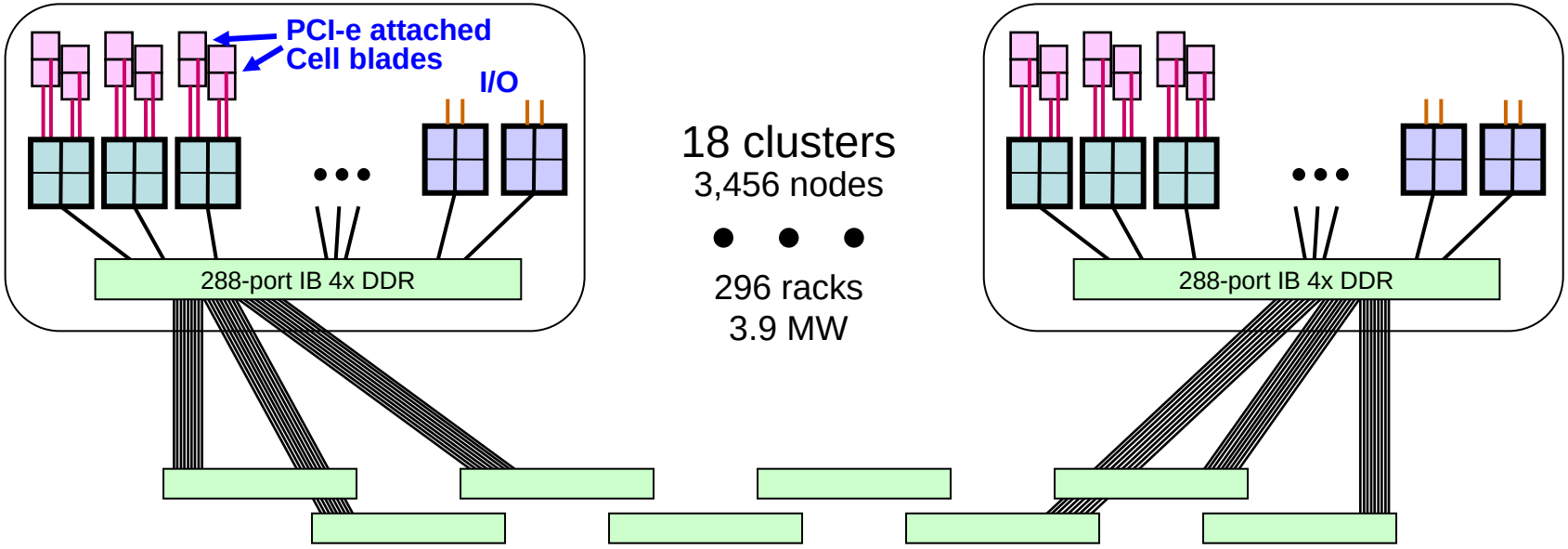
Applications on the Cell-accelerated Roadrunner machine

(Credit goes to many people working on the codes and doing performance measurements and modeling)

Roadrunner is a hybrid Cell-accelerated 1.4 PF system of modest size delivered in 2008

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

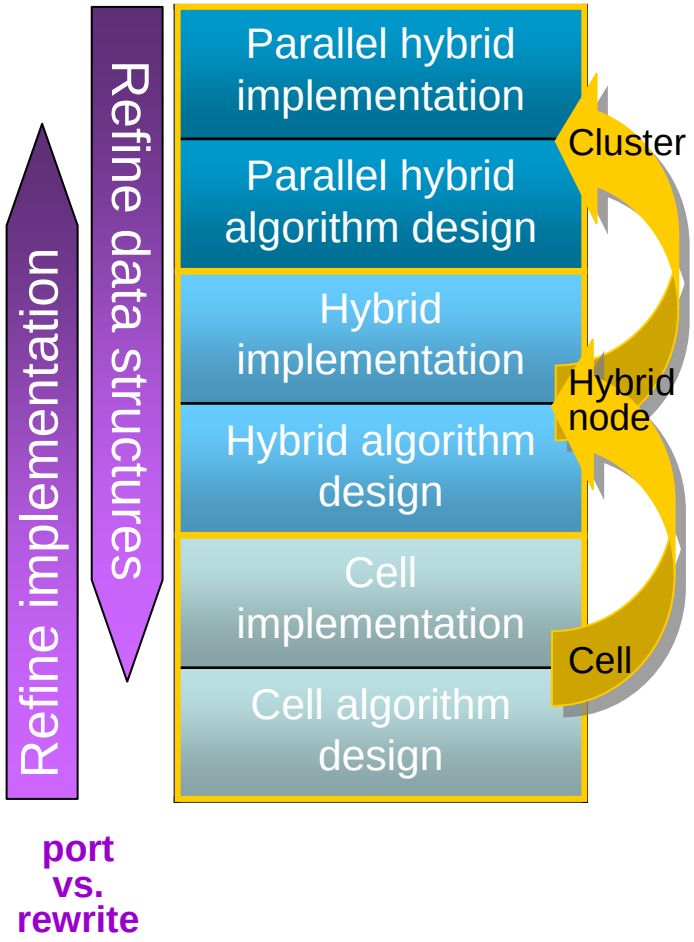
12,960 Cell eDP chips ⇒ 1.3 PF, 52 TB
6,912 dual-core Opterons ⇒ 50 TF, 28 TB



Eight 2nd-stage 288-port IB 4X DDR switches
12 links per CU to each of 8 switches

A few key algorithms are being targeted

- Radiation Transport
  - PARTISN (neutron transport via Sn)
    - Sweep3D (benchmark code)
  - MILAGRO (Implicit Monte-Carlo thermal)
- Particle methods
  - Molecular dynamics (SPaSM)
  - Particle-in-cell (VPIC)
- Eulerian hydro
  - Direct Numerical Simulation
- Linear algebra
  - LINPACK
  - Preconditioned Conjugate Gradient (PCG)

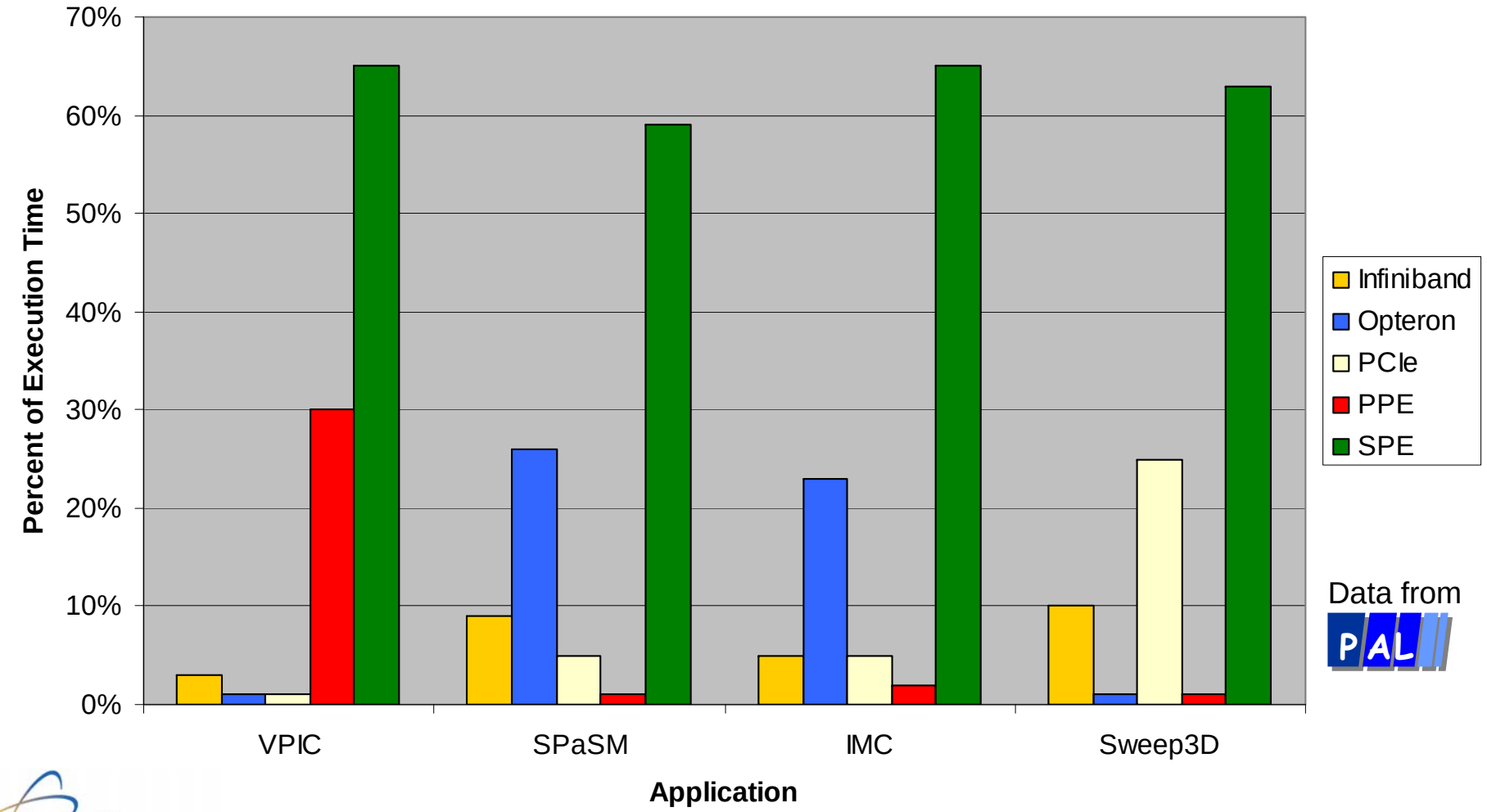


Cell and hybrid speedup results demonstrate clear success.

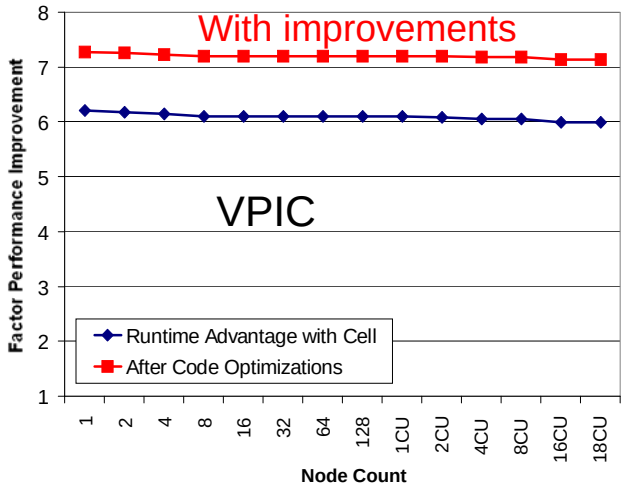
Application	Type	Class	Cell Only (kernels)		Hybrid
			Current	Roadrunner	(Current Cell + Infiniband to Opteron)
SPaSM	Science	full app	3x	4.5x	2.5x **
VPIC	Science	full app	9x	9x	6x
Milagro	Transport	full app	5x ##	6.5x ##	5x
Sweep3D	Transport	kernel	5x	9x	5x

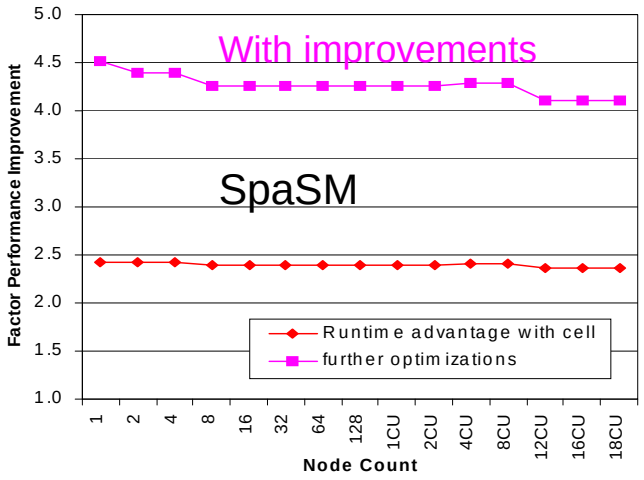
- all comparisons are to a single Opteron core
- parallel behavior unaffected, as will be shown in the scaling results
- ** Cell / hybrid SPaSM implementation does twice the work of Opteron-only code
- ## Milagro Cell-only results are preliminary

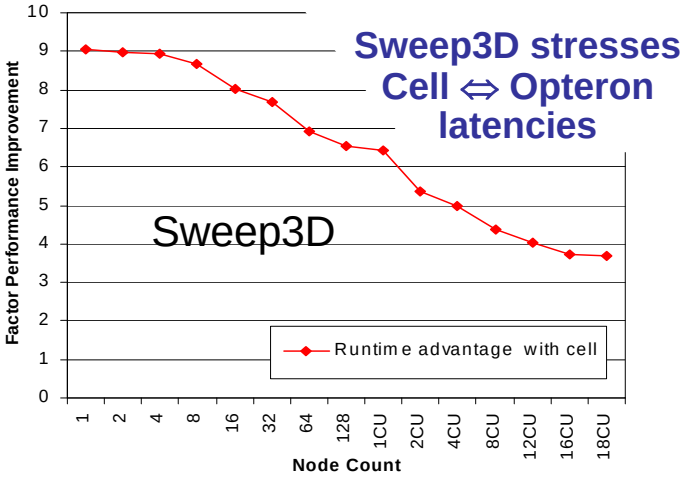
Roadrunner architecture is flexible - Applications are free to use hardware in most appropriate manner.

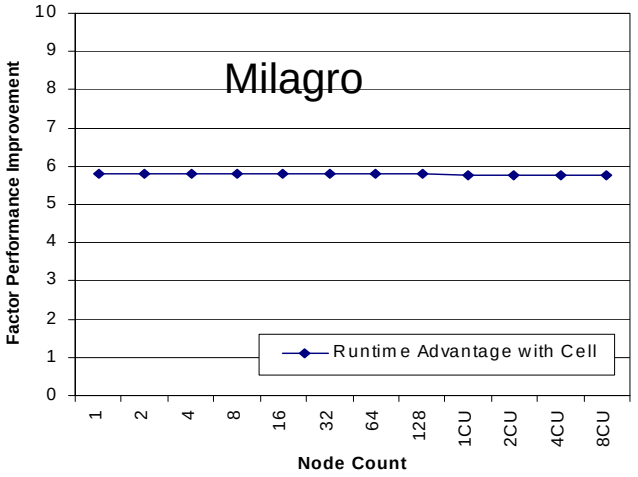


Our applications are predicted to scale out well on the final Cell-accelerated Roadrunner system









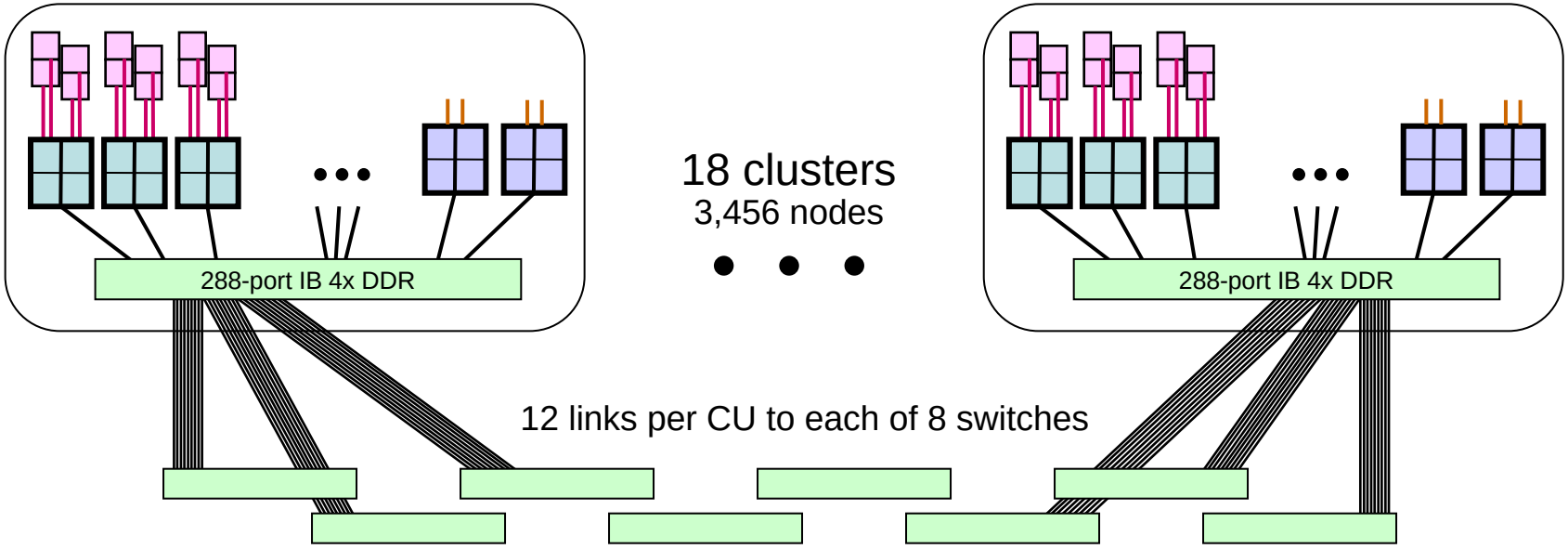
Data from LANL

Roadrunner at a Glance

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

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

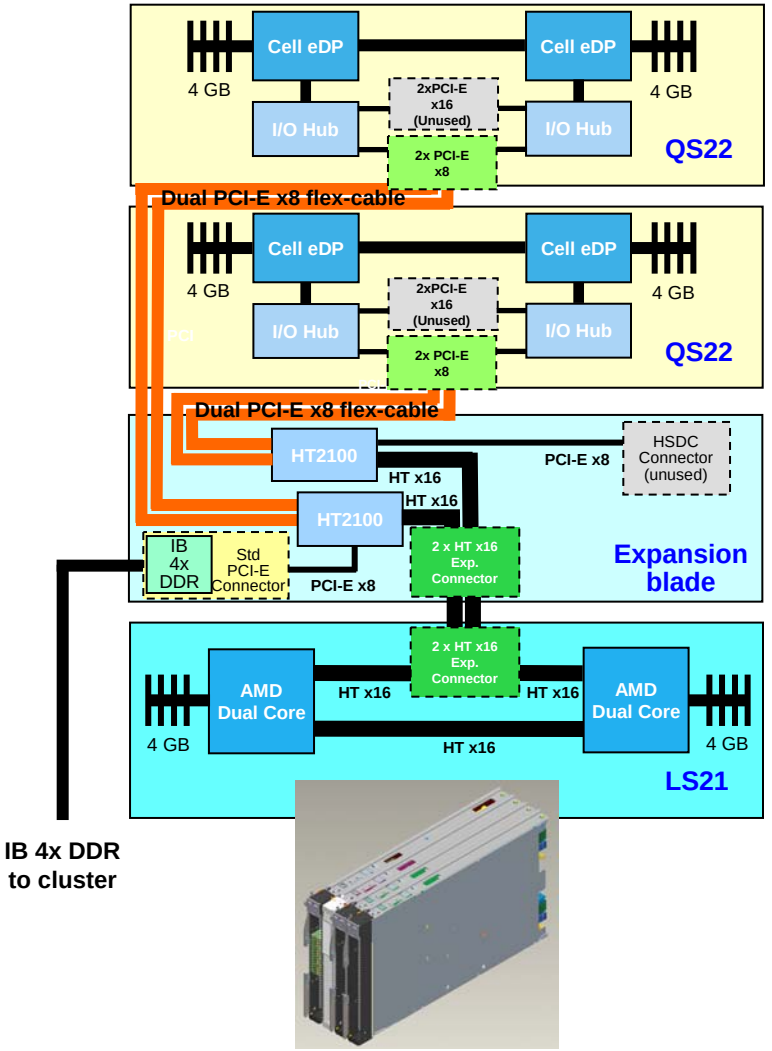
6,912 dual-core Opterons ⇒ 50 TF
12,960 Cell eDP chips ⇒ 1.3 PF



Eight 2nd-stage 288-port IB 4X DDR switches

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 internal PCI-E x8 links to LS21 and 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



More is information is available on the LANL Roadrunner home page

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

Roadrunner Architecture

Other Roadrunner talks

Computing Trends

Related Internet links

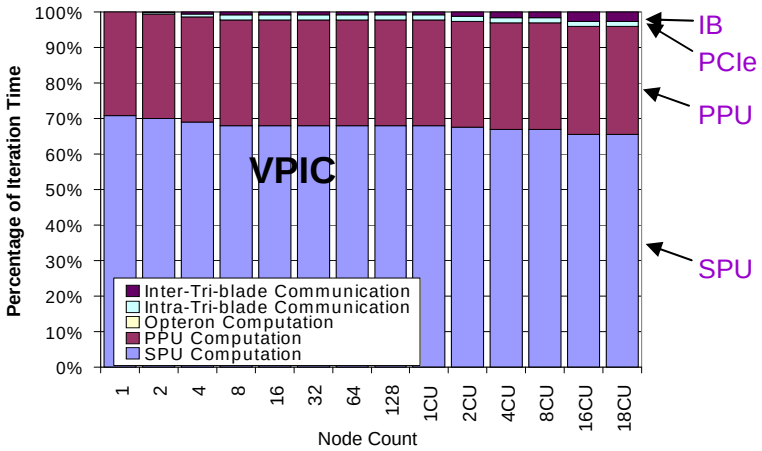
Extras

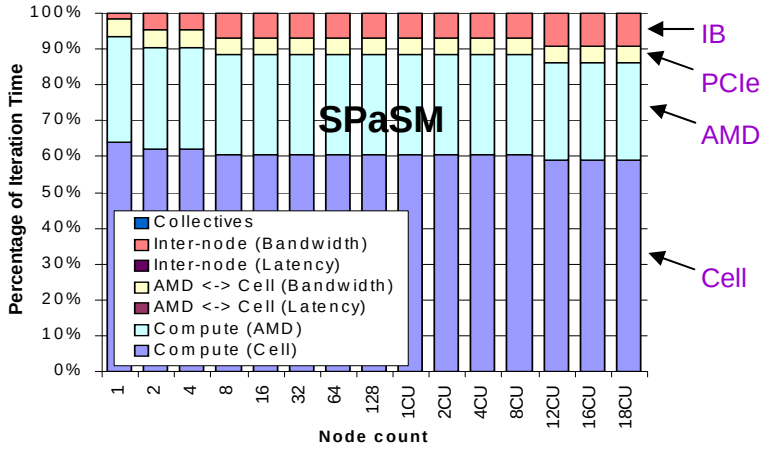
These results were achieved with a relatively modest level of effort.

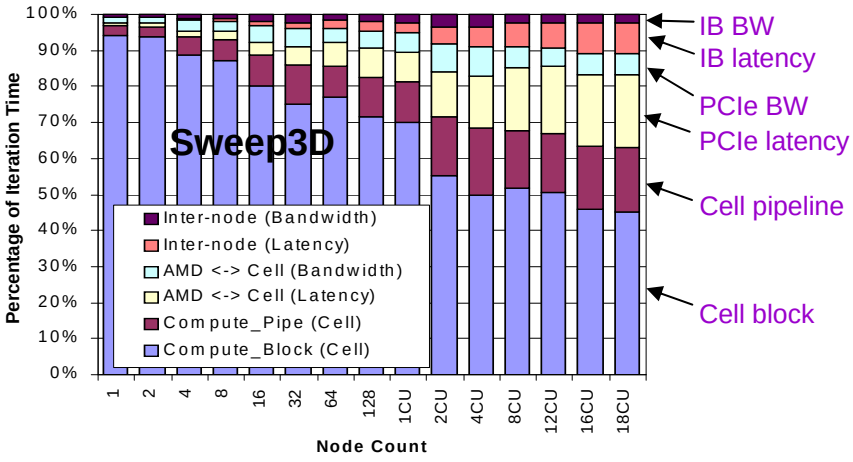
Code	Class	Language	Lines of code		FY07 FTEs
			Orig.	Modified	
VPIC	full app	C/C++	8.5k	10%	2
SPaSM	full app	C	34k	20%	2
Milagro	full app	C++	110k	30%	2 x 1
Sweep3D	kernel	C	3.5k	50%	2 x 1

- all staff started with little or no knowledge of Cell / hybrid programming
- 2 x 1 denotes separate efforts of roughly 1 FTE each
- most efforts also added code

The performance models give valuable break out information for further tuning & optimization



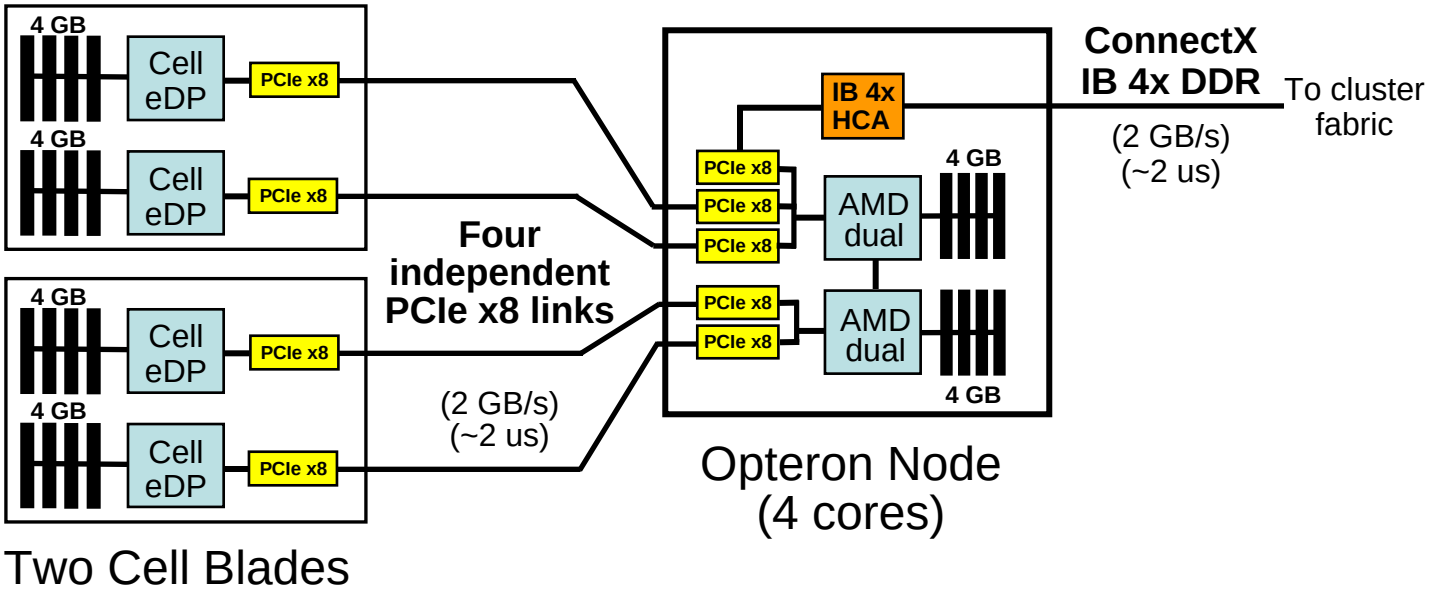




Data from

Roadrunner uses Cells to make nodes ~30x faster

400+ GFlop/s performance per hybrid node!



One Cell chip per Opteron core