IBM HPC DIRECTIONS

Dr Don Grice

ECMWF Workshop October 31, 2006

© 2006 IBM Corporation

What's Changing? The Rate of Frequency Improvement is Slowing

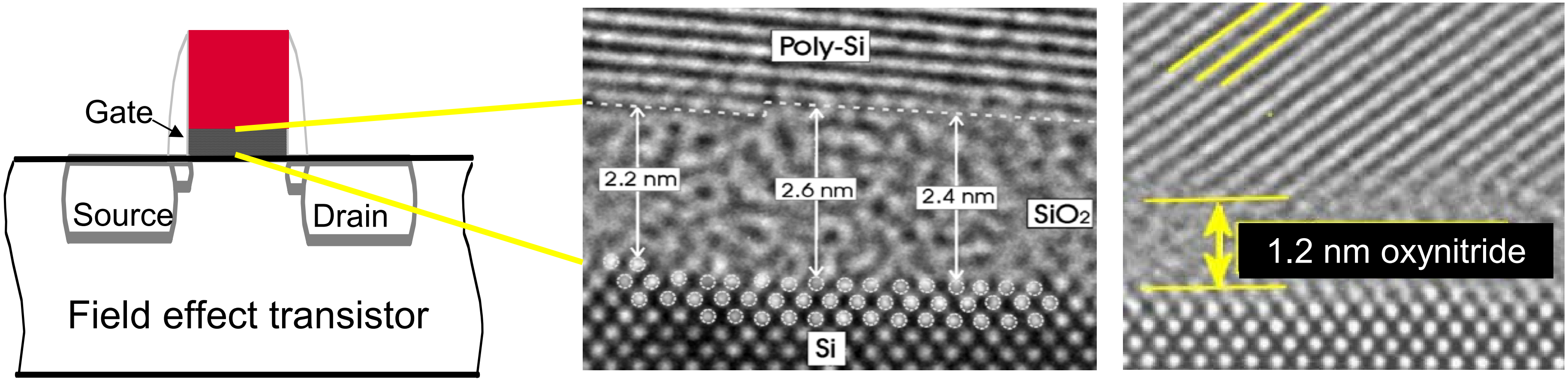
- Moore's Law (Frequency improvement) is a simplistic subset of Scaling
  - Circuit Density will continue to increase

BUT…

- Rate of Frequency Improvement is slowing
  - Leakage/Standby Power is increasing

- How do we adjust to the change in the rate of Frequency Improvement?
  - How do we deal with the Power and Power Density issues?

Why Scaling Breaks Down; We’re down to atoms



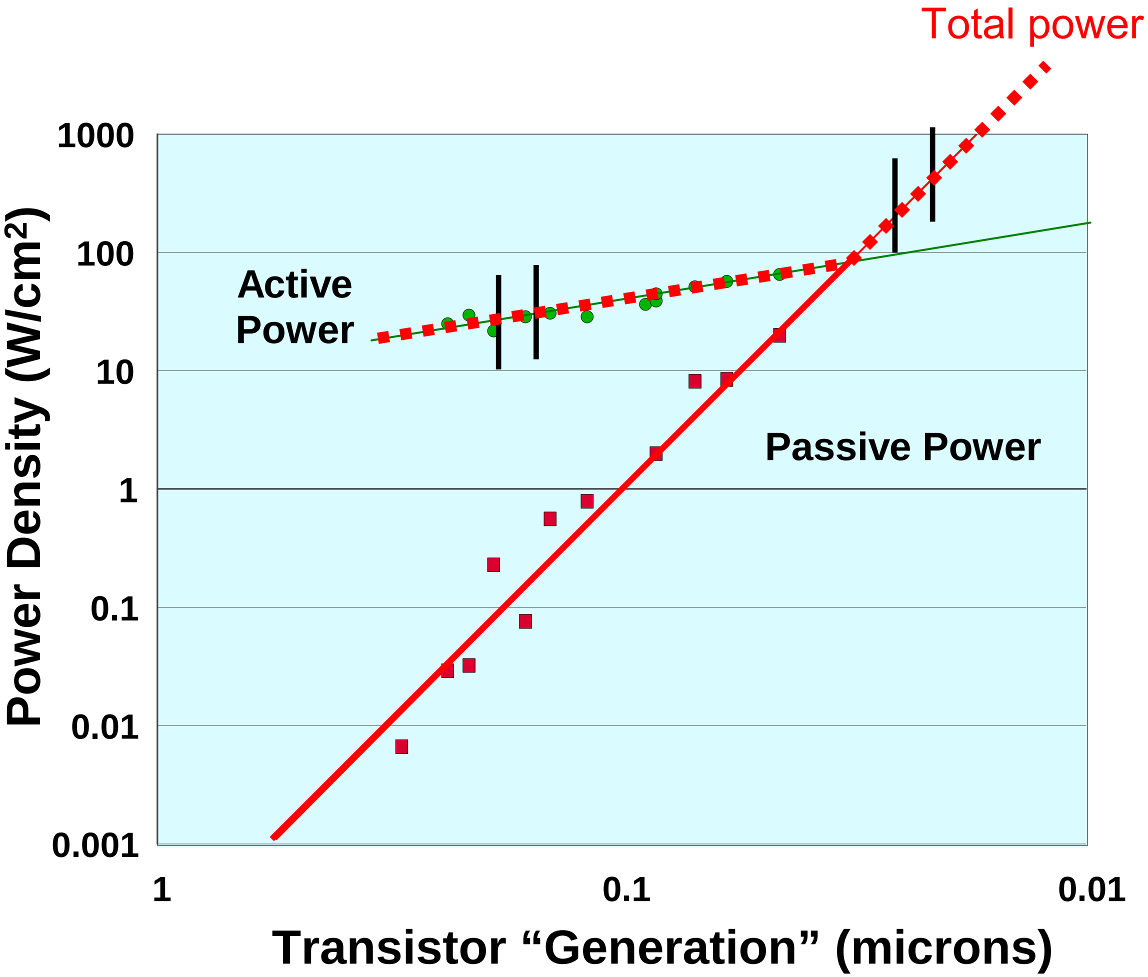
“Thick” gate oxide

Scaled gate oxide

- Consider the gate oxide in a CMOS transistor (the smallest dimensions today)
  - Assume only 1 atom high “defects” on each surrounding silicon layer
    - For a modern “scaled” oxide, 6 atoms thick, 33% variability is induced.
  - The bad news
    - Single atom defects can cause local current leakage 10-100x higher than average
  - The really bad news
    - Such “non-statistical behaviors” are appearing elsewhere in technology

Consider the Issue of Chip Power

- **Fundamental Changes**
  - **“Stopping” the chip no longer reduces chip power.**
  - **One must develop means to literally “unplug” unused circuits.**
  - **Software must become much more sophisticated to cope with selective shutdowns of processor assets.**
  - **Scaling produces profoundly different results when attempting to “push“ chip speeds**

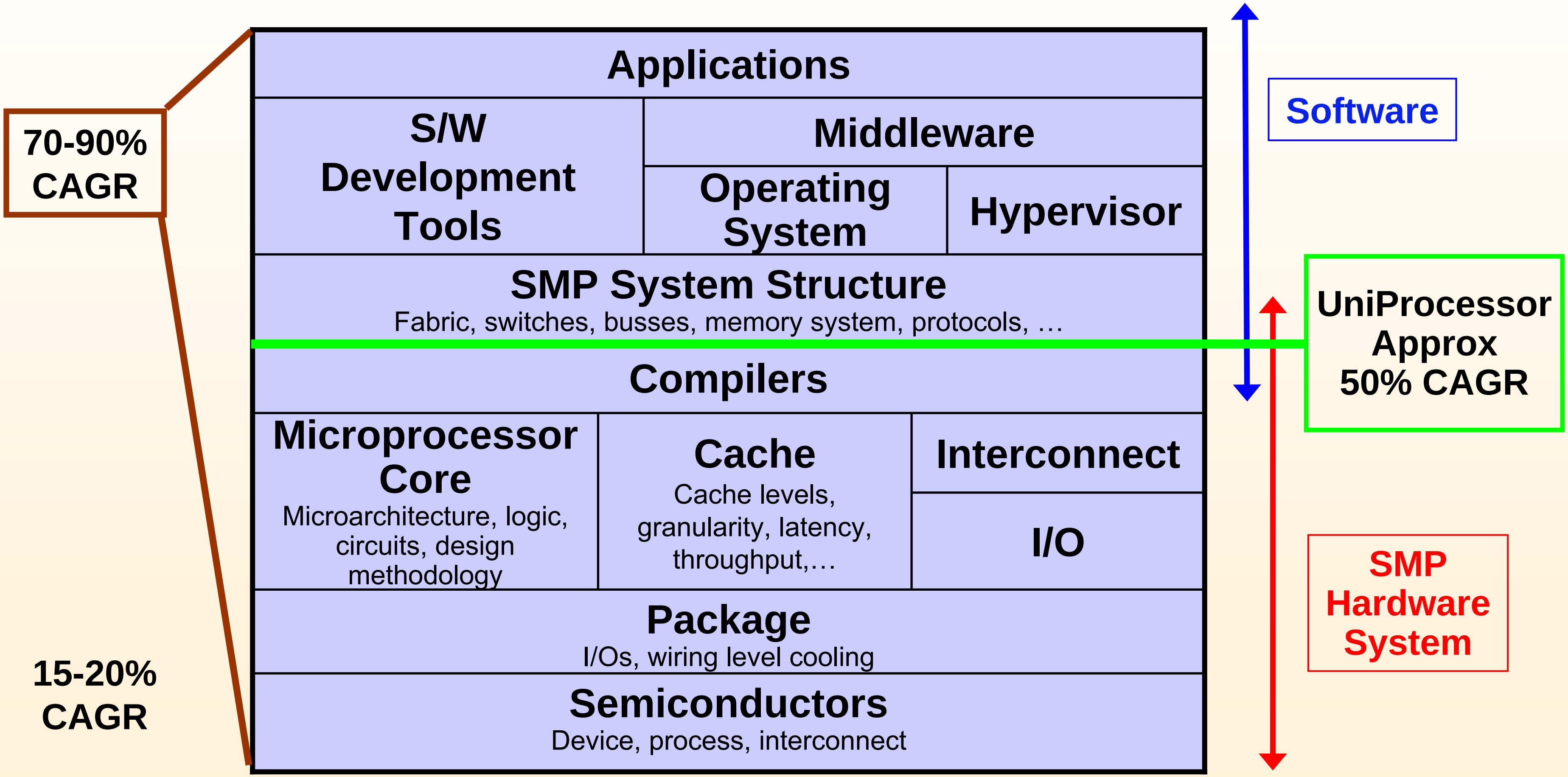


© 2006 IBM Corporation

System Performance Improvements

System performance gains of 70-90% CAGR derive from far more than semiconductor technology alone

Performance improvements will increasingly require system level optimization


70-90% CAGR
15-20% CAGR
Applications
S/W Development Tools
Middleware
Operating System
Hypervisor
SMP System Structure
Fabric, switches, busses, memory system, protocols, …
Compilers
Microprocessor Core
Microarchitecture, logic, circuits, design methodology
Cache
Cache levels, granularity, latency, throughput,…
Interconnect
I/O
Package
I/Os, wiring level cooling
Semiconductors
Device, process, interconnect
Software
UniProcessor Approx 50% CAGR
SMP Hardware System

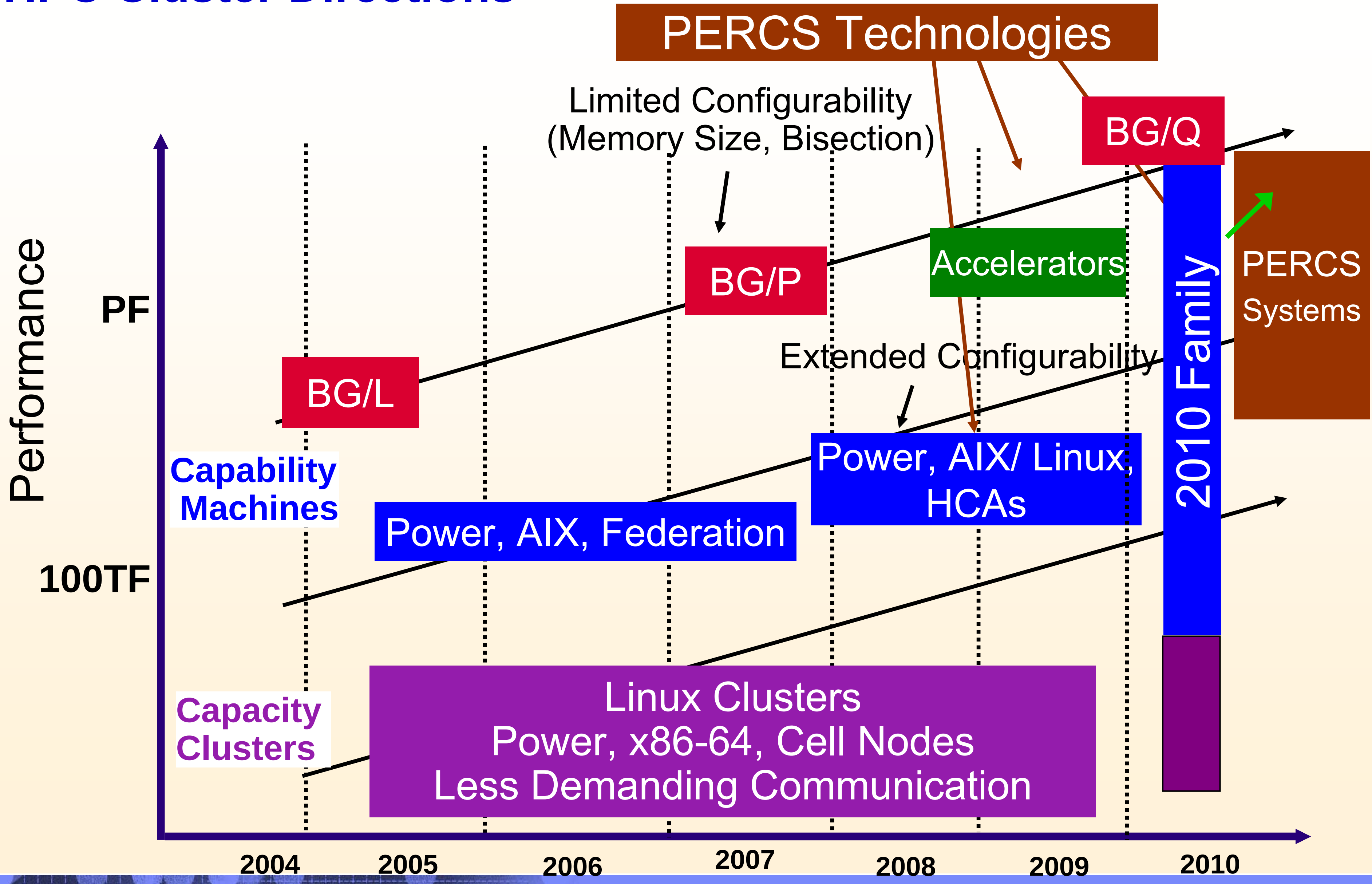
Multi-Core Options

- Homogeneous Symmetric Multi-Core General Purpose CPUs (Multiple Threads)
- Homogeneous Symmetric Multi-Cores with Specialized Instructions
- Heterogeneous Cores with Specialized Accelerators
- ‘System on a Chip’

Science Driven Design

© 2006 IBM Corporation

HPC Cluster Directions



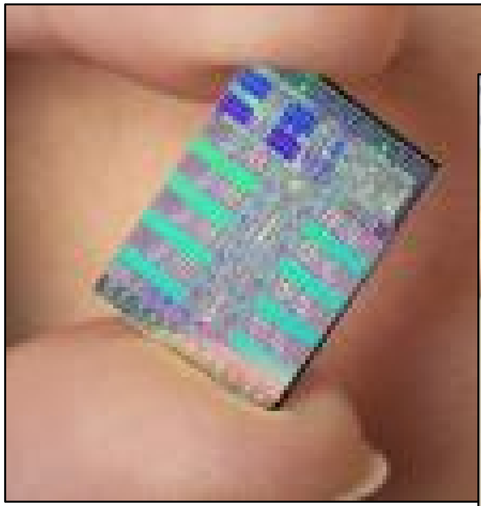
A Hybrid-Accelerator Machine ‘RoadRunner’

Roadrunner is a Critical Supercomputer Asset

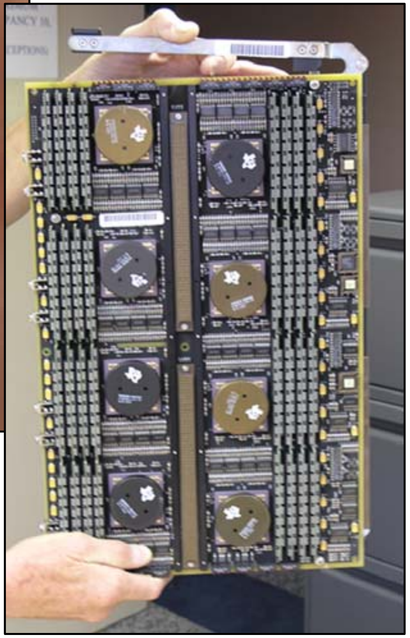
- Contract Awarded to September 8, 2006

- Critical component of stockpile stewardship
  - Initial system supports near-term mission deliverables
  - Hybrid final system achieves PetaFlops level of performance

- Accelerated vision of the future
  - Faster computation, not more processors



Cell processor (2007, ~100 GF)



CM-5 board (1994, 1 GF)

Roadrunner Goals

- Provide a large “capacity-mode” computing resource for LANL weapons simulations
  - Purchase in FY2006 and stand up quickly
  - Robust HPC architecture with known usability for LANL codes

- Possible upgrade to petascale-class hybrid “accelerated” architecture in a year or two
  - Follow future trends toward hybrid/heterogeneous computers
  - Capable of supporting future LANL weapons physics and system design workloads
  - Capable of achieving a **sustained** PetaFlop

Roadrunner System Overview

- IBM x3755 8-way Opteron servers as cluster nodes
- QLogic/PathScale & Voltaire InfiniBand (IB) cluster interconnects
- 144-node cluster size building blocks
- Vanilla Linux cluster software with diskless nodes

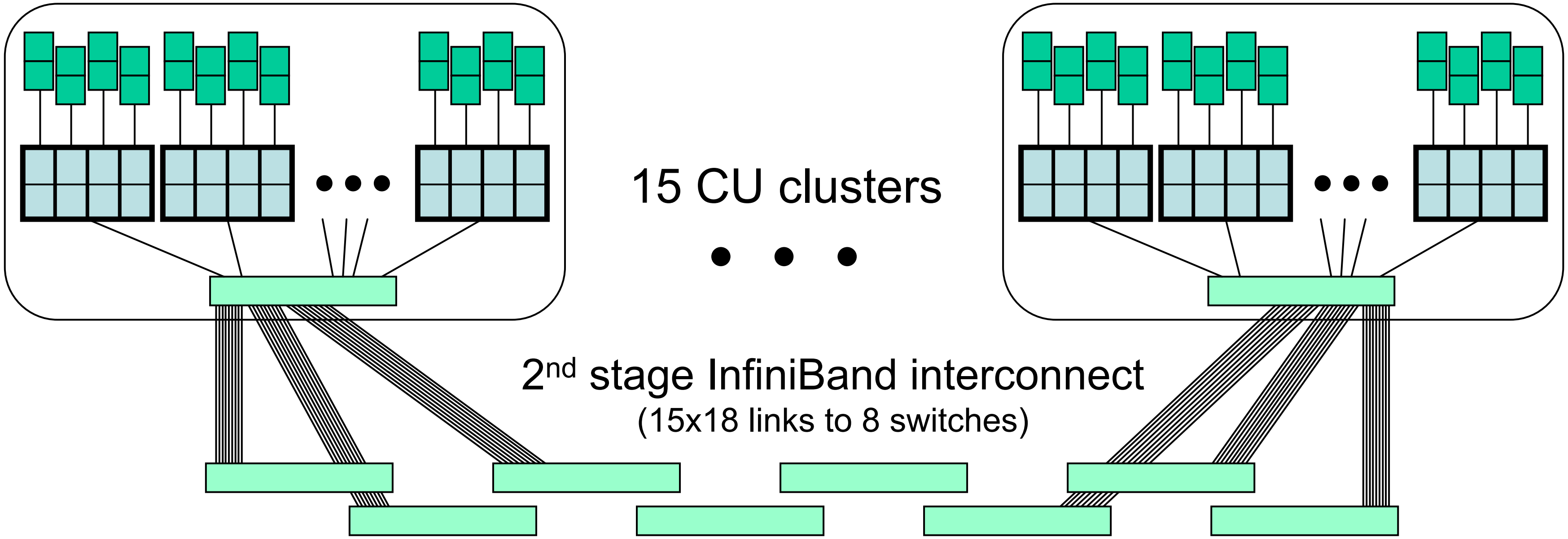
Final Classified System:

- 76 TF of Opteron
- IBM Cell Blades directly connected via IB to Opteron nodes for Accelerated

Accelerated Roadrunner

In aggregate:
8,640 dual-core Opterons + 16,560 eDP Cell chips
76 TeraFlops Opteron + ~1.7 PetaFlops Cell

"Connected Unit" cluster
144 quad-socket
dual-core nodes
(138 w/ 4 dual-Cell blades)
InfiniBand interconnects



15 CU clusters
● ● ●

2nd stage InfiniBand interconnect
(15x18 links to 8 switches)

Hybrid Programming

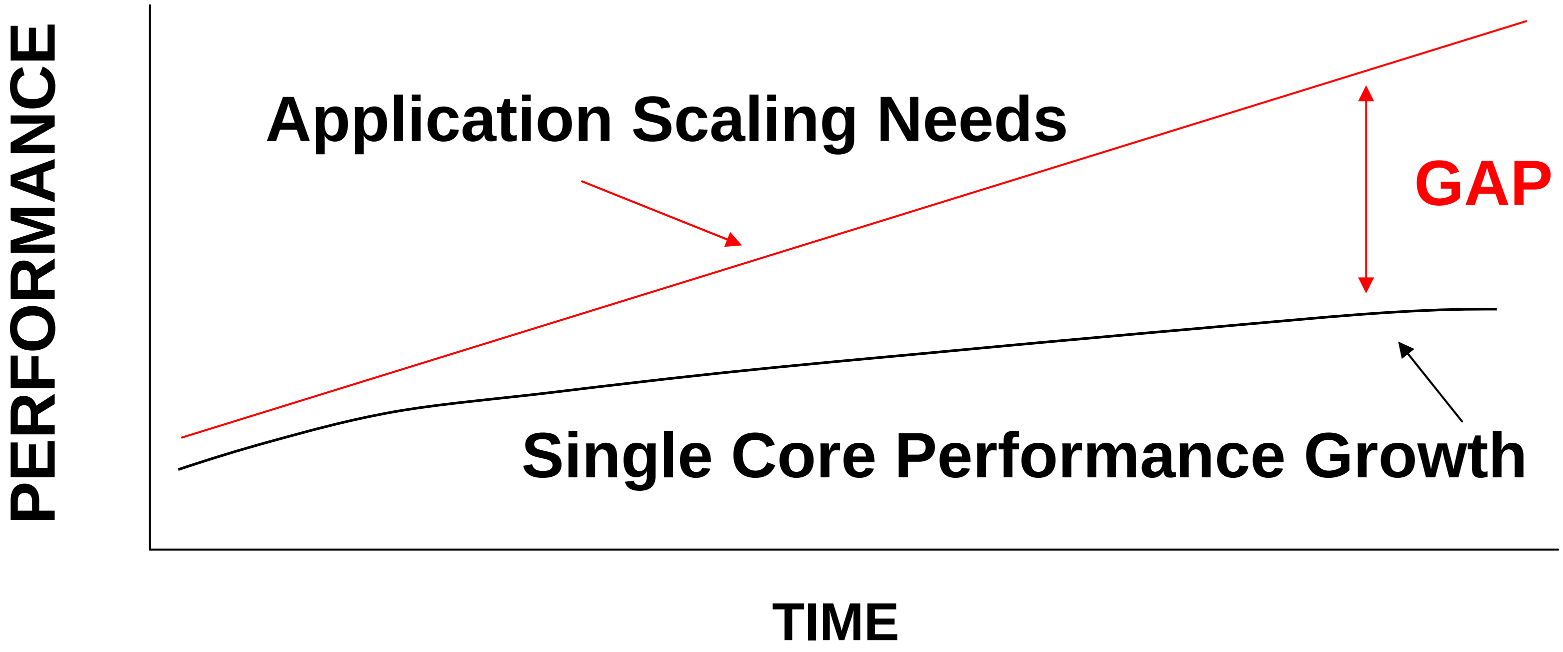
- Roadrunner is hybrid/heterogeneous
  - Standard Opteron-only parallel codes run unaltered on Roadrunner cluster nodes
  - Computationally intense kernels or entire modules or pieces are partially modified or rewritten to take advantage of Cells
    - Hopefully limit the source code impacted
- A hybrid code would have 3 distinct cooperating pieces
  1. Main code runs on Opteron of a node
  2. A Cell PPC code
  3. A Cell SPE code
  - Developer architects the cooperation now; tools may be able to help some in the future

Coping with the Decrease In the Rate of Frequency Improvement

PRODUCTIVITY TOOLS

© 2006 IBM Corporation

Application Performance Needs
vs
CPU Frequency Scaling



Key Problem: Frequency Improvements Do Not Match App Needs

Increasing Burden On The Application Design

Objective: Provide Tools to allow Scientists to Bridge the Gap

© 2005 IBM Corporation

BRIDGING THE GAP

Increased Parallelism (Petascale -> >100K CPUs)

- Need to Deal Application Scaling Issues
- Initial Application Design
- Application Tuning
- Application Debug

Improved Performance Efficiency

- Algorithmic Changes
- Overlapped Communication
- Improved Synchronization

Specialized Cores/Accelerators

- Library and Language Support

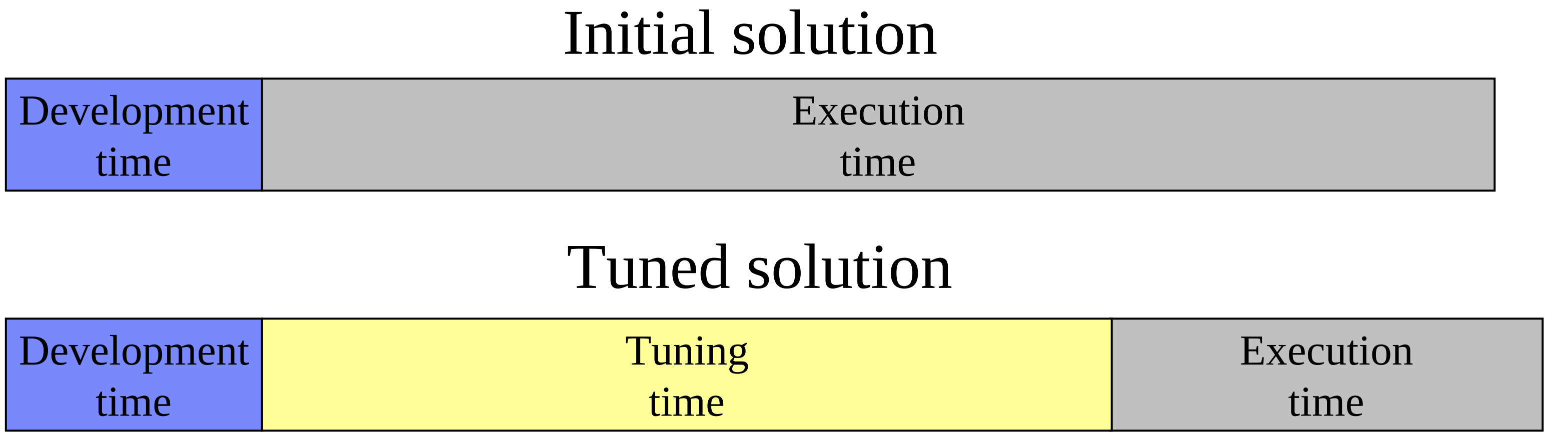
© 2005 IBM Corporation

STEPS IN APPLICATION DEPLOYMENT

- DEVELOPMENT
- TUNING/DEBUG
- EXECUTION (N-TIMES)

NEED TO IMPROVE EFFICIENCY IN ALL 3

Performance Tuning Dilemma



Initial solution

Development time

Execution time

Tuned solution

Development time

Tuning time

Execution time

© 2005 IBM Corporation

Programmer Productivity

- **Developing new parallel applications**
  - IDE environments
  - Libraries
- **Debugging parallel applications**
  - Tracing and parsing hooks
  - Static analysis tools
  - Rational
- **Performance tuning parallel applications**
  - HPC Toolkit, FDPR, compiler feedback, static analysis tools
- **Application/system tuning parameters**
  - CPO

© 2005 IBM Corporation

Ideas to address sustained performance

- **Overlap computation/communication:**
  - RDMA Exploitation
- **Collective Communication overheads:**
  - Collective Offload Engines
- **Communication latency:**
  - Cache injection
- **Memory latency:**
  - Pre-fetch hints to subsystems and applications (SMT assist threads)
- **Algorithmic**
  - Better tools to assist in smarter application development

    (Locking & Message passing impacts)

Programming Models and Library Support

- **Models supported today**
  - MPI – Two sided: Message Passing Interface
  - LAPI – One sided programming model
  - SHMEM – shared memory programming model
- **Possible Future Models**
  - X10 concepts being applied to HPLS initiatives
  - UPC – Unified Parallel C
- **Component Libraries**
  - ESSL, PESSL enhancements, COIN-OR

IBM will support all emerging models that see wide HPC adoption

© 2005 IBM Corporation

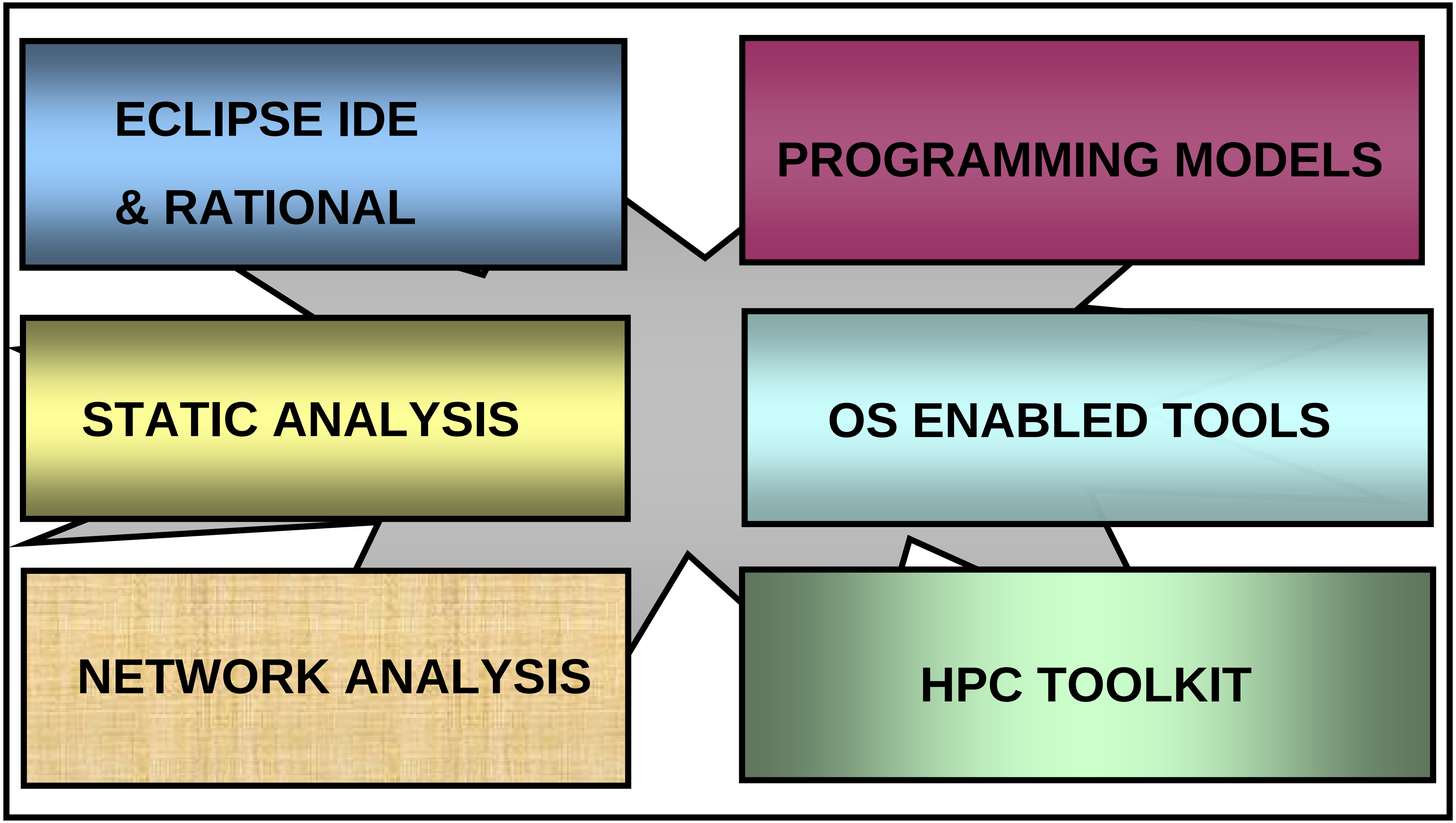
Compilers

IBM compiler technology is critical to reducing time to solution:

- *Productivity* focus
  - Support new programming concepts, environments and languages as they achieve increased adoption (UPC, X10-HPLS, etc.)
  - Enhance existing environments (MPI, OpenMP, Fortran, C/C++, Java)
  - Support for CPO technologies
  - Static Analysis tools for MPI applications
- *Performance* focus
  - Dynamic compilation
  - Memory wall
  - SMT exploitation
  - Support new hardware-software concepts (e.g. pseudo vector morph)
  - Drive towards zero cache-miss execution of performance critical paths
  - Exploit automatic RDMA overlap

© 2005 IBM Corporation

Productivity Solution Vision



ECLIPSE IDE & RATIONAL

PROGRAMMING MODELS

STATIC ANALYSIS

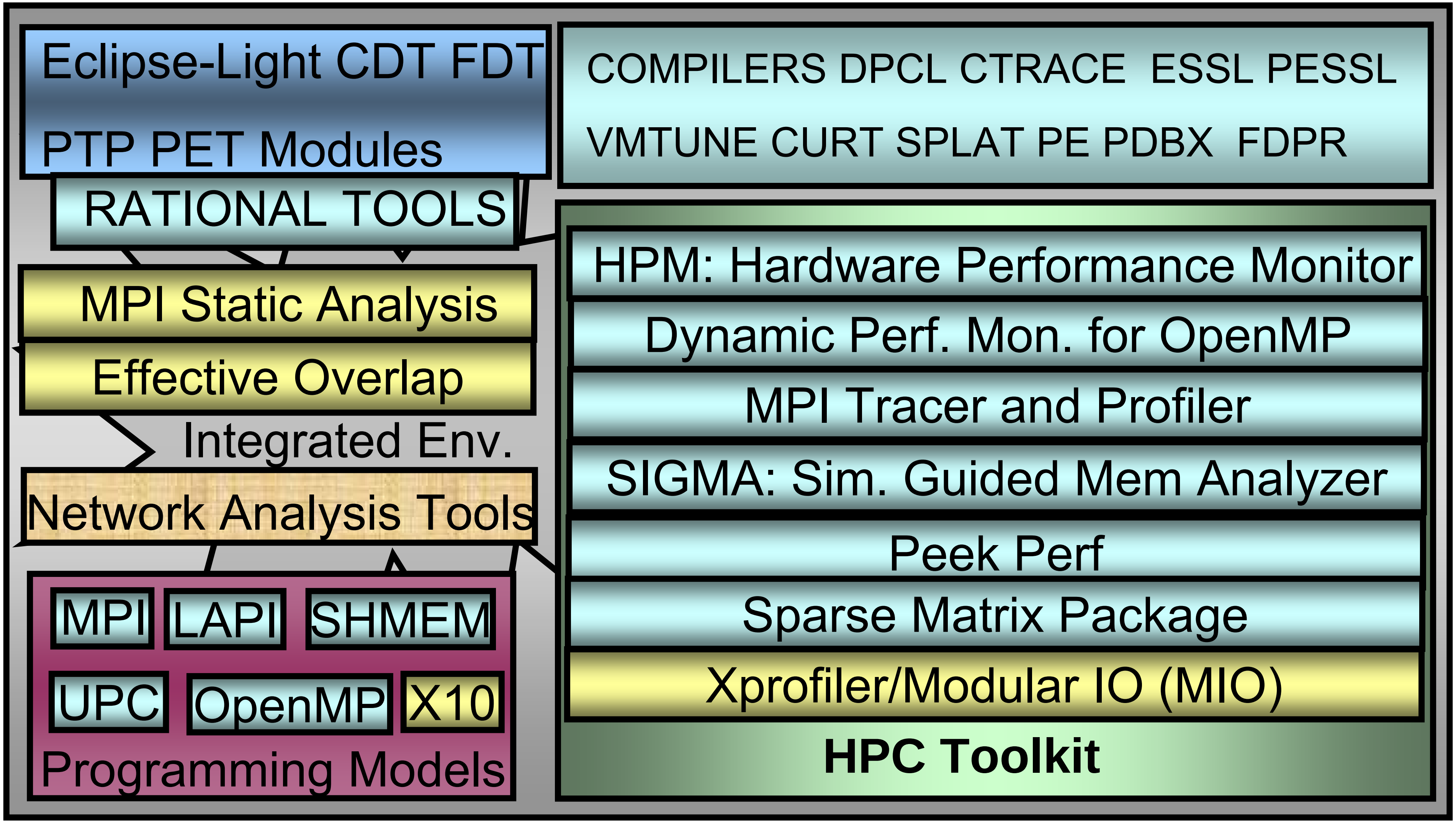
OS ENABLED TOOLS

NETWORK ANALYSIS

HPC TOOLKIT

© 2005 IBM Corporation

Overall PERCS Productivity Solution Vision



Eclipse-Light CDT FDT

PTP PET Modules

RATIONAL TOOLS

MPI Static Analysis

Effective Overlap

Integrated Env.

Network Analysis Tools

MPI LAPI SHMEM

UPC OpenMP X10

Programming Models

COMPILERS DPCL CTRACE ESSL PESSL

VMTUNE CURT SPLAT PE PDBX FDPR

HPM: Hardware Performance Monitor

Dynamic Perf. Mon. for OpenMP

MPI Tracer and Profiler

SIGMA: Sim. Guided Mem Analyzer

Peek Perf

Sparse Matrix Package

Xprofiler/Modular IO (MIO)

HPC Toolkit

© 2005 IBM Corporation