SOS13
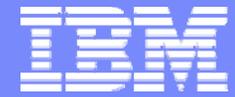## 13th Workshop on Distributed Supercomputing

What will be your company's HPC successes in 2012?

March, 2009

IBM Systems
Simplify your IT.

IBM HPC DIRECTIONS

Dr Don Grice

SOS 13 - 3/09

Technology Trends and Challenges

There are 3 major challenges on the road to and beyond Petascale computing:

- Supplying and paying for the power required to run the machines which can be equivalent to a small town!

- Reliability, Availability, and Serviceability because of the law of large numbers and Soft Errors

- Finding efficient ways to program machines that will support MILLIONS of simultaneous processes!

IBM Ultra Scale Approaches

- Blue Gene – Maximize Flops Per Watt
  with Homogeneous Cores
  by reducing Single Thread Performance

- Power/PERCS – Maximize Single Thread Performance
  with Homogeneous Cores

- Roadrunner – Use Heterogeneous Cores and
  an Accelerator Software Model
  to Maximize Flops Per Watt
  and keep High Single Thread Performance

Roadrunner: Science, Cell and a Petaflop/s

International Supercomputing Conference

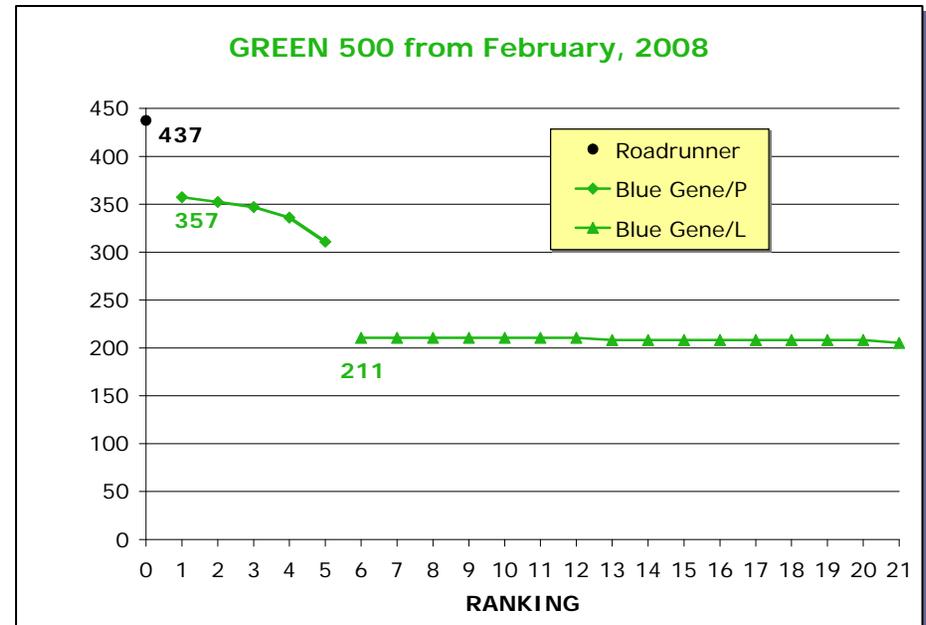## Special Roadrunner Session, 6-18-2008

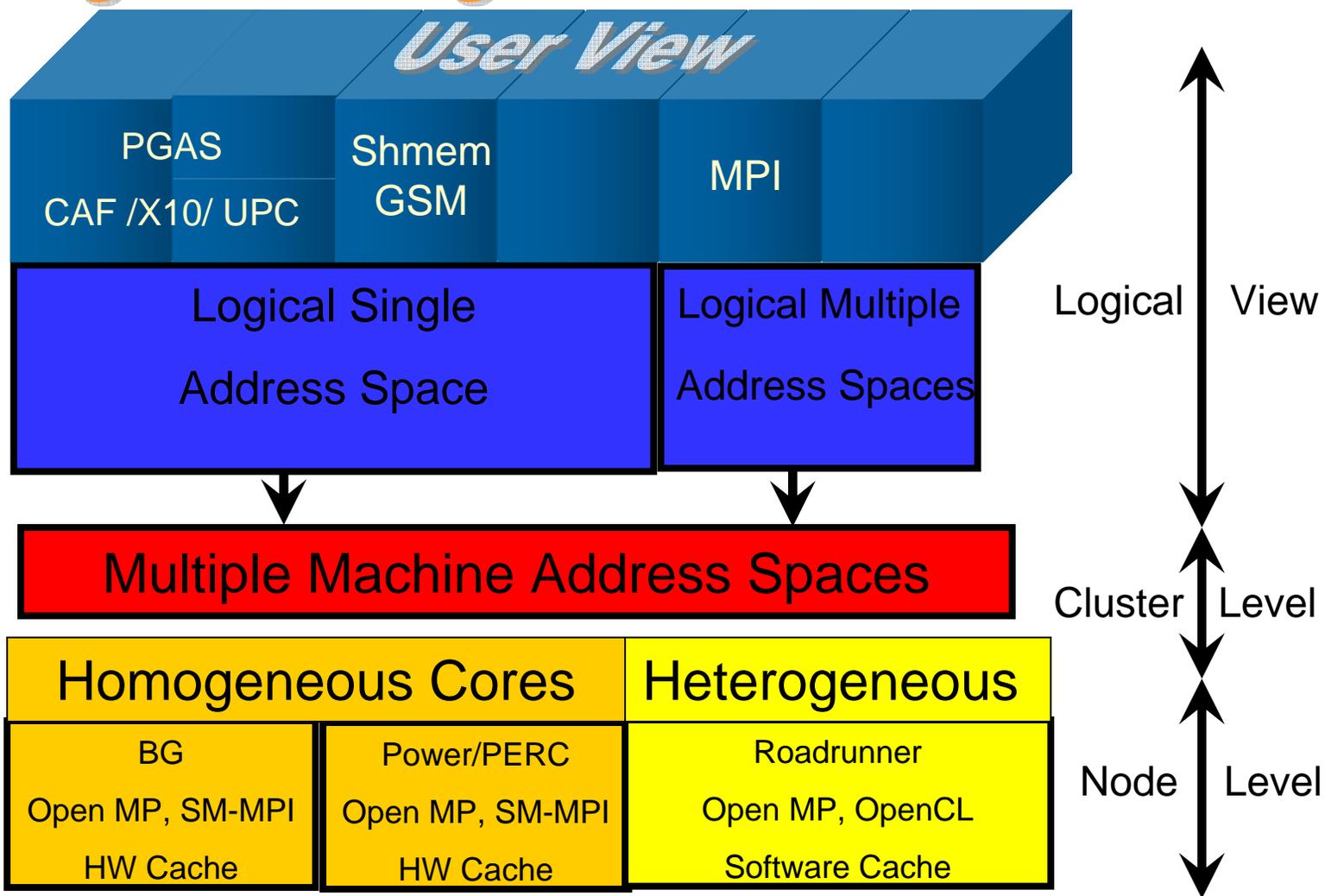Andy White

Los Alamos

Don Grice

IBM

Achieving a petaflop/s in less than 3 days demonstrates the stability of the Roadrunner system.

- **Full system available**
  - 8:30 am, Friday, May 23

- **Full system job launch tests begin**
  - 3:00 pm, Friday, May 23

- **First full system LINPACK launch**
  - 8:30 pm, Friday, May 23
  - Node failure after running an hour

- **Successful LINPACK runs**
  - 5:45 pm, Saturday, May 24 (879 TF/s)
  - 2:45 pm, Sunday, May 25 (945 TF/s)
  - 1:10 am, Monday, May 26 (997 TF/s)
  - 3:30 am, Monday, May 26 (Petaflop/s)

GREEN 500 from February, 2008

Legend:
- Roadrunner
- Blue Gene/P
- Blue Gene/L

(chart values noted: 437, 357, 211; axis label RANKING)

Roadrunner is also very energy efficient, 437 MF/s per watt.

Los Alamos
NATIONAL LABORATORY
EST. 1943

IBM®

Operated by Los Alamos National Security, LLC for NNSA

NNSA

Programming Models: Architecture

User View

PGAS	Shmem	MPI
CAF /X10/ UPC	GSM	

Logical Single Address Space	Logical Multiple Address Spaces

Multiple Machine Address Spaces

Homogeneous Cores		Heterogeneous
BG	Power/PERC	Roadrunner
Open MP, SM-MPI	Open MP, SM-MPI	Open MP, OpenCL
HW Cache	HW Cache	Software Cache

Logical | View

Cluster | Level

Node | Level

Design for Availability and Serviceability

- HW Failures will happen

- Machines need to be designed to keep them from effecting user applications as much as possible

- Repairing failures quickly and easily is critical

- Application Impact:
  - With enough circuits error rates will not be 'zero'
  - Need algorithms that are fault tolerant
    - Compute more than once – or more than one way?
    - Internal Consistency checking?

Things to Think About
Algorithm Changes for New System Limits?

- Computing will be 'free' but Pin BW will be expensive
    - Memory BW
    - Communication BW

- Need to restructure algorithms around the new limits

- Roadrunner Linpack used Redundant computing
  to reduce Communication BW
- DGEMM was restructured to minimize Memory BW

- Do we need a new FFT?

Panel Questions

Hypothesis 1

With users' relentless appetite for HPC we could expect systems with 100 PFlops peak soon after ~2012.

- **What type of systems do you think will first achieve this?**
  - **Power efficient, optimized Heterogeneous systems**
- **Will they be general-purpose?**
  - **General purpose in the sense of 'programmable' but probably targeted in terms of memory sizes, SMP sizes, etc. (i.e. not MDGrape like)**
- **How much electrical power will they consume?**
  - **20-60MW depending on the Hybrid mixture**

What will be the standard way of programming applications for these systems?
- **There will be advancements in more abstract languages like APGAS, Logical Shared Memory, etc.**
- **These languages need to then feed down into the heterogeneous run time environments.**

Hypothesis 2

The HPC industry faces huge challenges including mounting power consumption, maintaining system availability with increasing component volumes, decreasing memory bandwidth per core, software to scale to millions of processes. Yet vendors seem able to re-invent solutions on a regular basis.

- **What paradigm shifts, if any, do you see occurring by 2012?**
  - **There will be shifts toward Heterogeneous Cores and many cores**
  - **There needs to be a software shift to raise the level of abstraction so that the application folks are not swamped by the need to scale to the extremely large number of cores we will be using**
  - **New system limits need to be addressed (RAS, Pin BW vs Compute)**
- **Where might you look for new partners?**
  - **There will be an increased sense of open source community as new language constructs and 'middle ware' (OpenCL/OpenMPI etc) take on the challenge of heterogeneous cores.**

Hypothesis 3.

**The current world economy is drastically different to any ever seen
in the lifetime of the HPC industry.
Credit may effectively disappear, funds may become more centrally controlled,
hyperinflation arise from capital injections, and markets shrink.
Yet HPC users benefit from healthy competition sustained
by the current market size.**

- **What do you think is your company's best strategy for survival (in HPC) ?**
    - **Reduce the number of offerings and options in the high end by focusing on Heterogeneous System Configurations.**
    - **Converging the number of offerings will reduce the flexibility of high end offerings but will greatly reduce the development expenses.**
- **How do you think HPC customers can realistically help you?**
    - **Focus requests for 'differences in offerings' to those features that are needed for strategic deployments, as opposed to 'the need to be different for funding purposes'.**

THANK YOU