Platform™

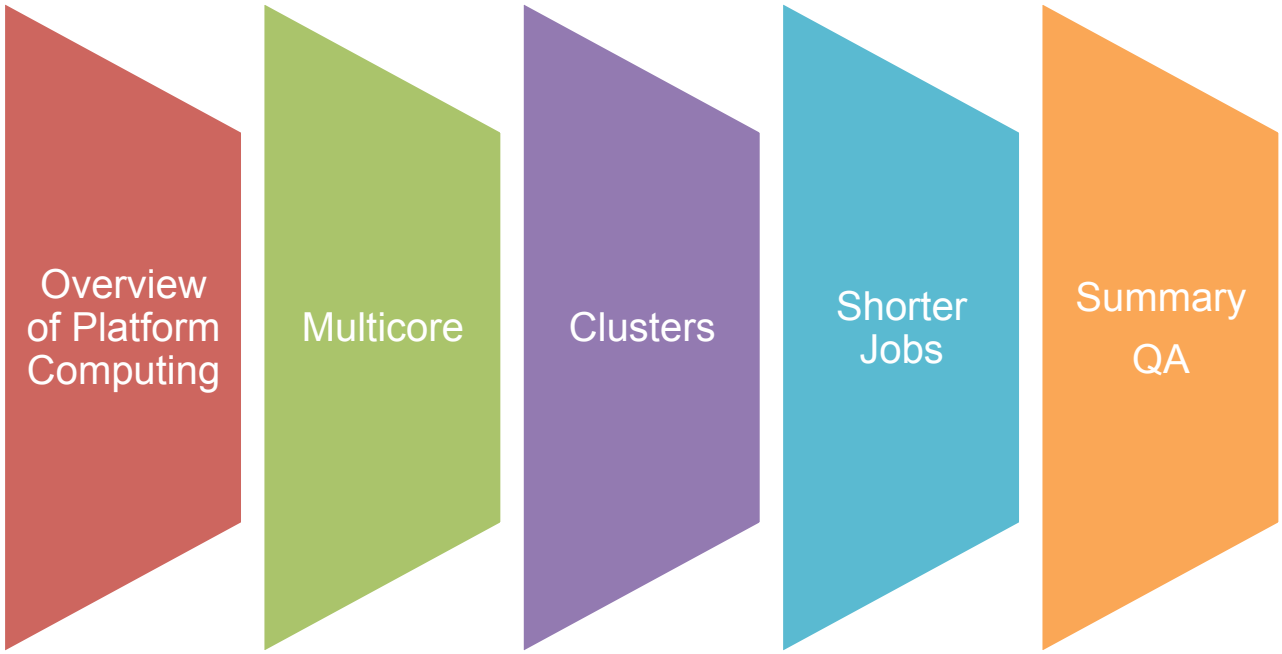
Powering High Performance

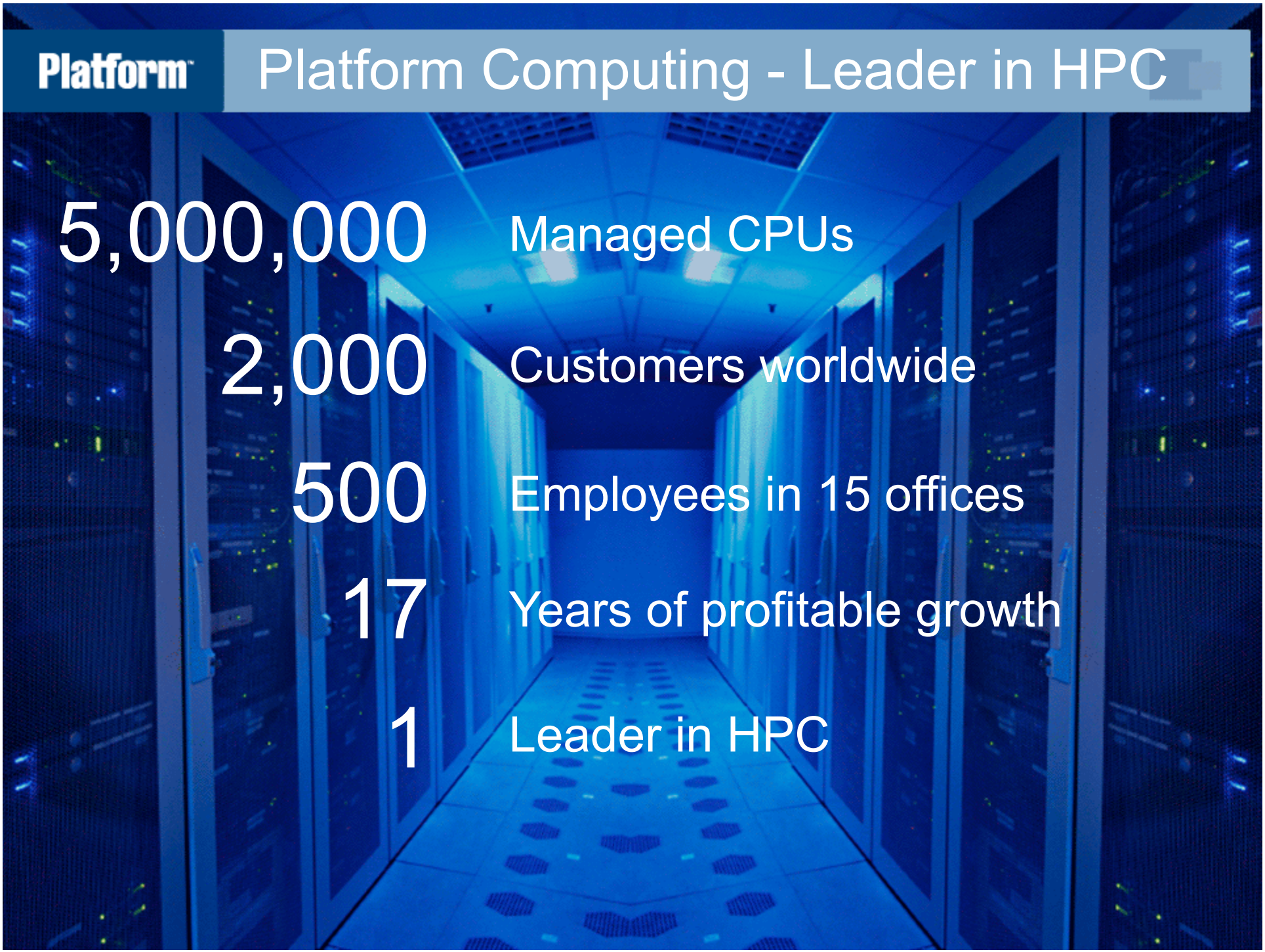
Trends in HPC

Presenter: Robert Stober
Date: May 2009

Platform

Agenda

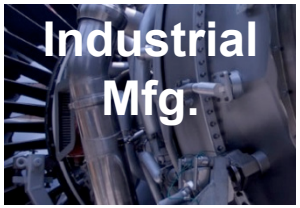



Platform
Platform Computing - Leader in HPC
5,000,000
Managed CPUs
2,000
Customers worldwide
500
Employees in 15 offices
17
Years of profitable growth
1
Leader in HPC

Platform Industries Served by Platform

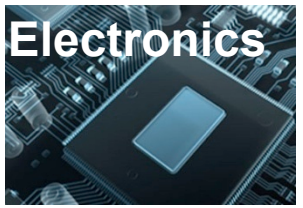
- CERN
- DoD, US
- DoE, US
- ENEA
- Georgia Tech
- Harvard Medical School
- Japan Atomic Energy Inst.
- MaxPlanck Inst.
- MIT
- Shanghai SC
- Stanford Medical
- TACC
- U. Of Georgia
- U. Tokyo
- Washington U.

- BNP
- Citigroup
- Fortis
- HSBC
- KBC Financial
- JPMC
- Lehman Brothers
- LBBW
- Mass Mutual
- MUFG
- Nomura
- Prudential
- Sal. Oppenheim
- Société Générale



- Airbus
- BAE Systems
- Boeing
- Bombardier
- Deere & Company
- Ericsson
- Honda
- General Electric
- General Motors
- Goodrich
- Lockheed Martin
- Nissan
- Northrop Grumman
- Pratt & Whitney
- Toyota
- Volkswagen

- Agip
- BP
- British Gas
- China Petroleum
- ConocoPhillips
- EMGS
- Gaz de France
- Hess
- Kuwait Oil
- PetroBras
- Petro Canada
- PetroChina
- Shell
- StatoilHydro
- Total
- Woodside



- AMD
- ARM
- Broadcom
- Cadence
- Cisco
- Infineon
- MediaTek
- Motorola
- NVidia
- Qualcomm
- Samsung
- Sony
- ST Micro
- Synopsys
- TI
- Toshiba

- Abott Labs
- AstraZeneca
- Celera
- DuPont
- Eli Lilly
- Johnson & Johnson
- Merck
- National Institutes of Health
- Novartis
- Partners Health Network
- Pharsight
- Pfizer
- Sanger Institute

Other Industries

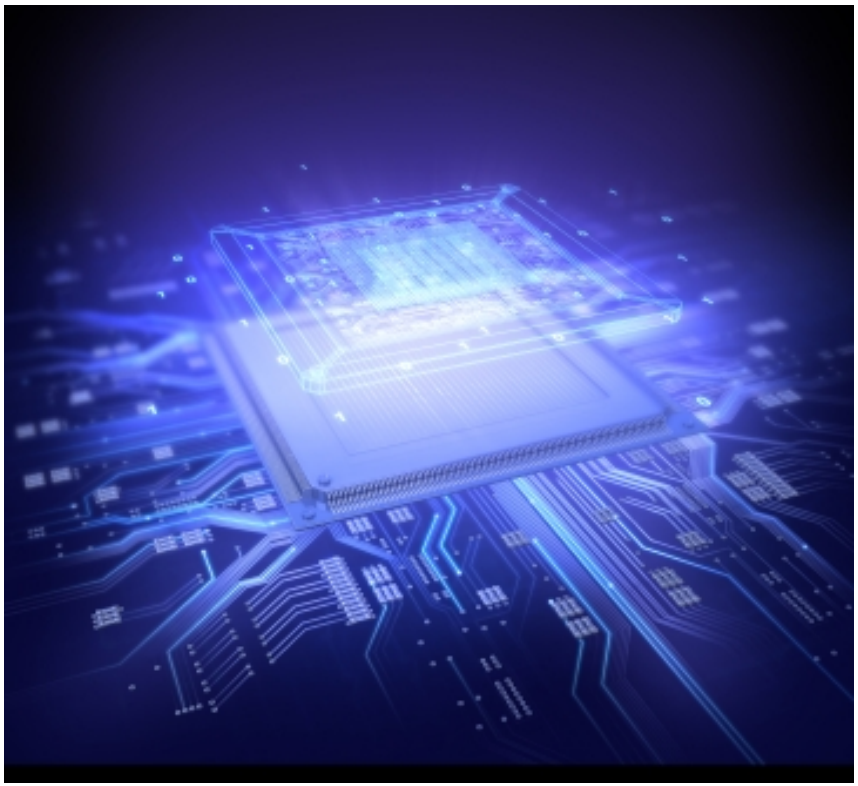
AT&T Bell Canada Cingular DreamWorks Animation SKG GE

IRI Telecom Italia Telefonica Walt Disney Co.

Platform Cluster Manager (PCM)

- PCM used to be called OCS
- PCM is a fully integrated, end-to-end solution including a complete range of tools necessary to simply deploy, run and manage an HPC cluster.
- Platform PCM is now available CX1
- Platform LSF has been available on the larger systems for some time.

The Trend Toward Multicore



- Processor Granularity
- Prior versions of Platform LSF allocated jobs at the processor granularity.
- Platform LSF can now be configured to consider processors, cores or threads as job slots. This is a cluster-wide configuration parameter

```
# set in lsf.conf
EGO_DEFINE_NCPUS=cores
```

Job Binding

- The kernel may not give optimal job performance
- It may place too many job processes on the same processor or core
- Or it may load balance processes from a hot cache to a cold cache
- Platform LSF can be configured to bind jobs to processors, cores, or threads

Job Binding

- Platform LSF processor binding provides hard processor binding functionality for sequential LSF jobs
- For parallel jobs, Platform LSF binds the job at the first execution host, not other remote hosts
- Processor binding can be configured on the application or cluster level
- Limitation: Processor binding is supported on hosts running Linux with kernel version 2.6 or higher.

Job Binding

- BIND_JOB=BALANCE policy instructs Platform LSF to balance the job across the available cores.
- The BIND_JOB=PACK policy directs Platform LSF to bind the job to a single processor
- The binding policy can also be delegated to the user through the BIND_JOB=USER and BIND_JOB=USER_CPU_LIST policies.

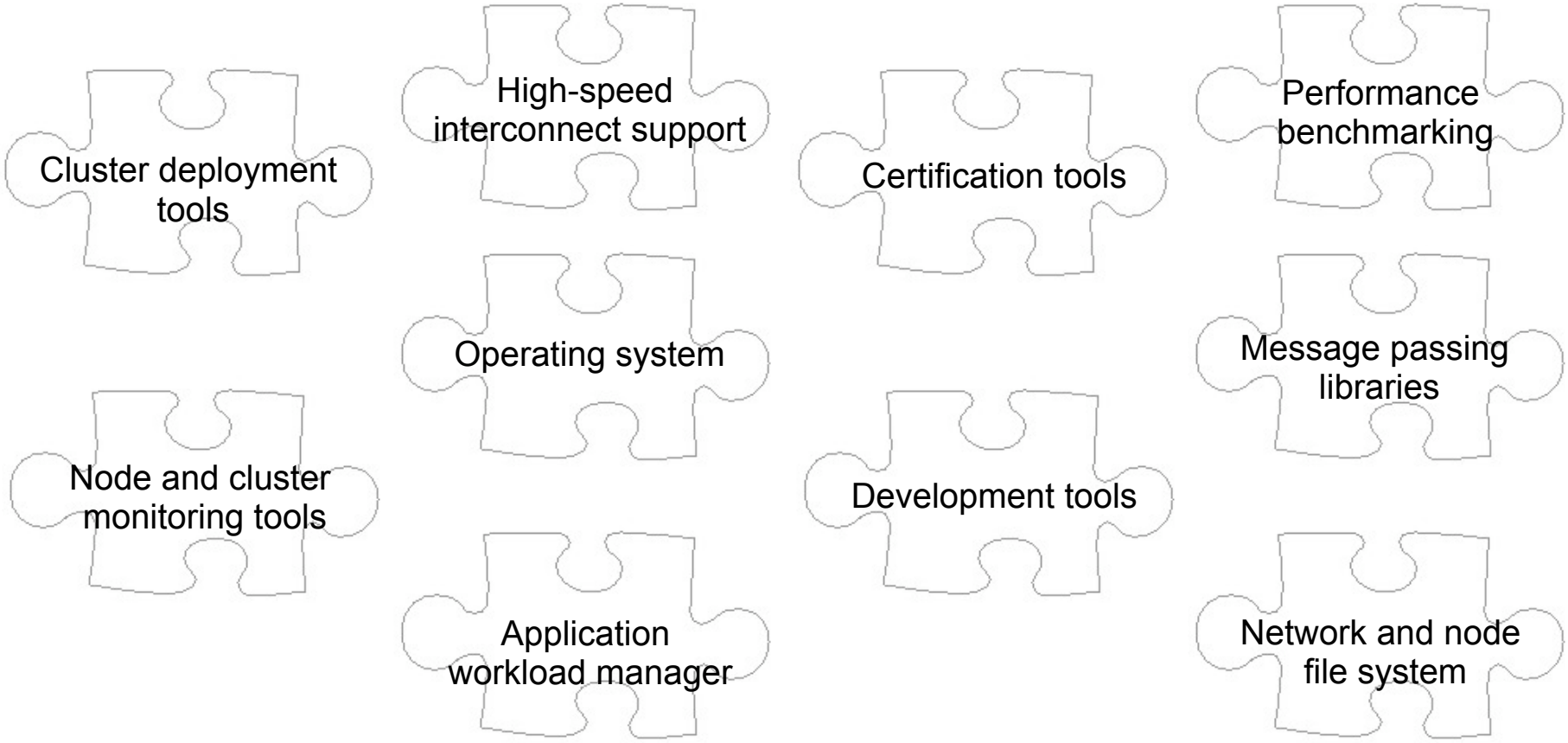
Platform

The Trend Towards HPC

- Organizations are constantly trying solve bigger problems, and many are turning to HPC to solve them.
  - Low cost operating system
  - Scalable
  - Open Source software infrastructure
  - Optional high speed interconnect and/or parallel file system
  - High value, low perceived cost

Platform Building a Cluster is Complicated

- It's a Jigsaw puzzle…



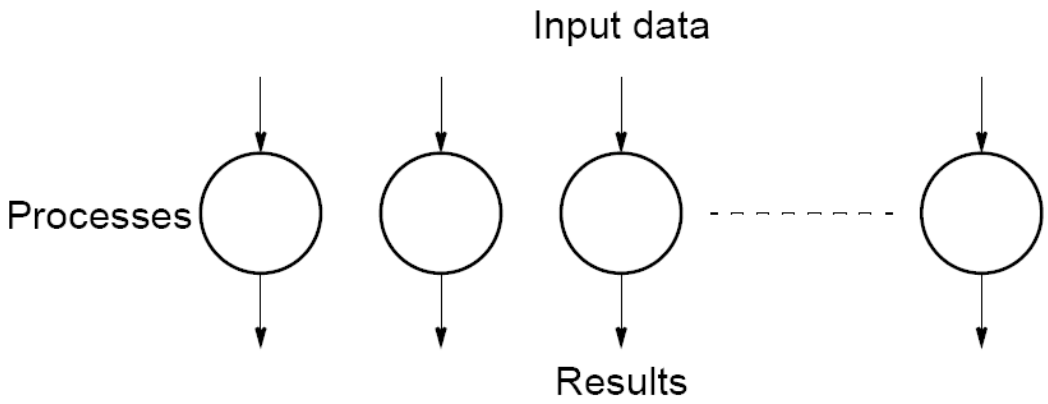
Need to *integrate* multiple products and tools from multiple sources

Platform Cluster Manager (PCM)

- PCM used to be called OCS
- PCM is a fully integrated, end-to-end solution including a complete range of tools necessary to simply deploy, run and manage an HPC cluster.
- Platform PCM is now available CX1
- Platform LSF has been available on the larger systems for some time.

Embarrassing Parallel Jobs

- A clear trend in many industries is that job volumes have been increasing while job run-times have been getting shorter.
- Many of these are embarrassingly parallel



No communication or very little communication between processes
Each process can do its tasks without any interaction with other processes

Platform

Embarrassing Parallel Jobs

An **embarrassingly parallel** workload (or *embarrassingly parallel problem*) is one for which little or no effort is required to separate the problem into a number of parallel tasks. This is often the case where there exists no dependency (or communication) between those parallel tasks. (Wikipedia)

Platform

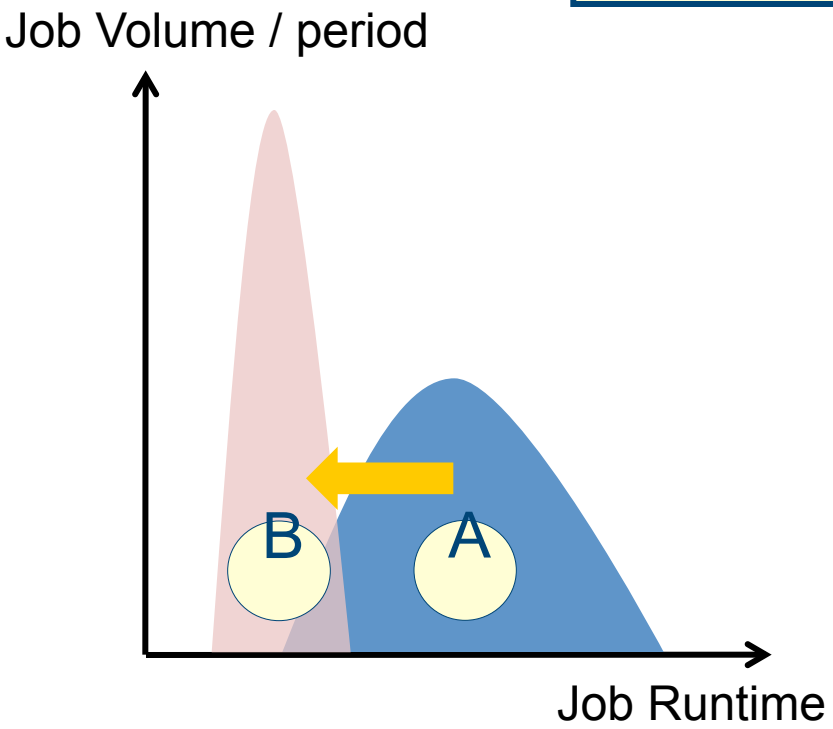
Embarrassing Parallel Jobs

- Design of Experiments (DoE) techniques in mechanical engineering a model may be run repeatedly with different inputs
- Stochastic analysis in financial modeling - Portfolio value may be computed repeatedly based on a range of randomized inputs
- Electronic device verification and regression - Semiconductor modeling based on an exhaustive set of initial starting conditions
- Image Processing - Rendering a sequence of frames, or searching for a pattern match in a set of existing images.
- Pharmaceutical research - Modeling the interaction of a candidate drug with particular protein targets

Embarrassing Parallel Jobs

- **In some industries, job volumes & cluster capacities are increasing, while job durations are simultaneously decreasing.**

Even with no increase in job volumes, shorter run-times and larger multi-CPU / multi-core clusters result in *dramatic load increases* on the scheduler!



Case "A"
- 1,000 cores
- Ave job run time 10 minutes
- # of jobs 1,000,000

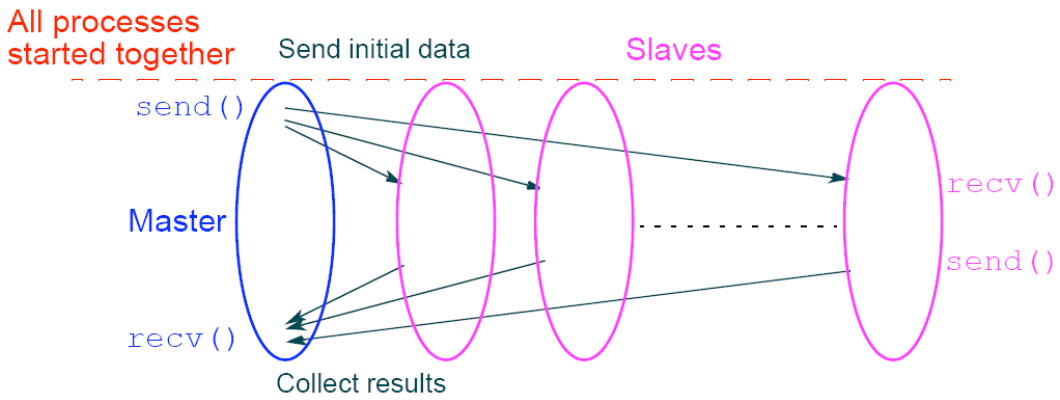
Scheduler handles **~ 6,000 jobs / hour**

Case "B"
- 4,000 cores
- Ave job run time 2 minutes
- # of jobs 1,000,000

Scheduler handles **~ 120,000 jobs / hour**

MPI as Job Scheduler

- Workload managers typically allocate the requested number of execution nodes and start the job on the first node
- Some applications developers are using MPI to schedule the jobs onto the nodes



Usual MPI approach

Platform

MPI as Job Scheduler

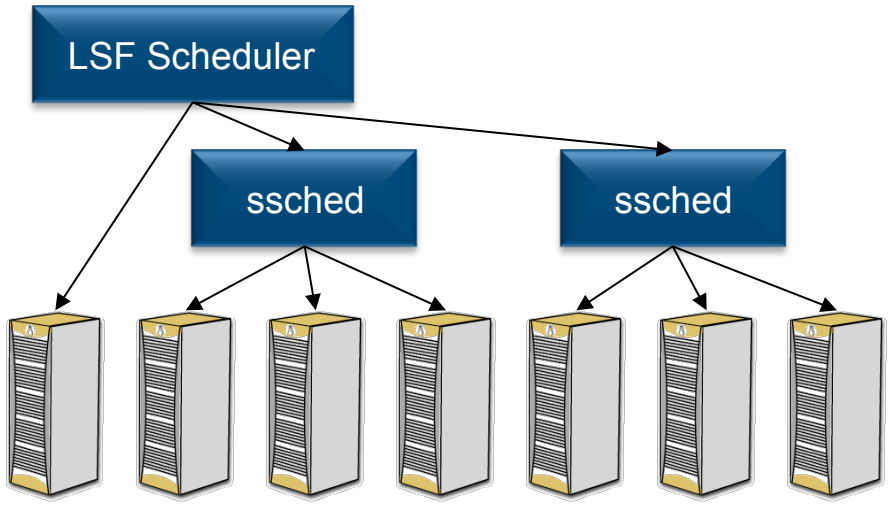
- MPI does not have the capability to handle fault tolerance
- The (adhoc) MPI scheduler is not dynamically scalable
- There's no task-level accounting
- Overhead may be considerably higher
- Costs $ to build and maintain

LSF Session Scheduler

- The new session scheduler supports dramatic increases in job throughput allowing large volumes of jobs to be managed as tasks on pre-allocated machines
  - Higher throughput / lower latency
  - Superior management of related tasks
  - Supports > 50,000 tasks / per user
  - two-tier scheduling – preserves existing job semantics

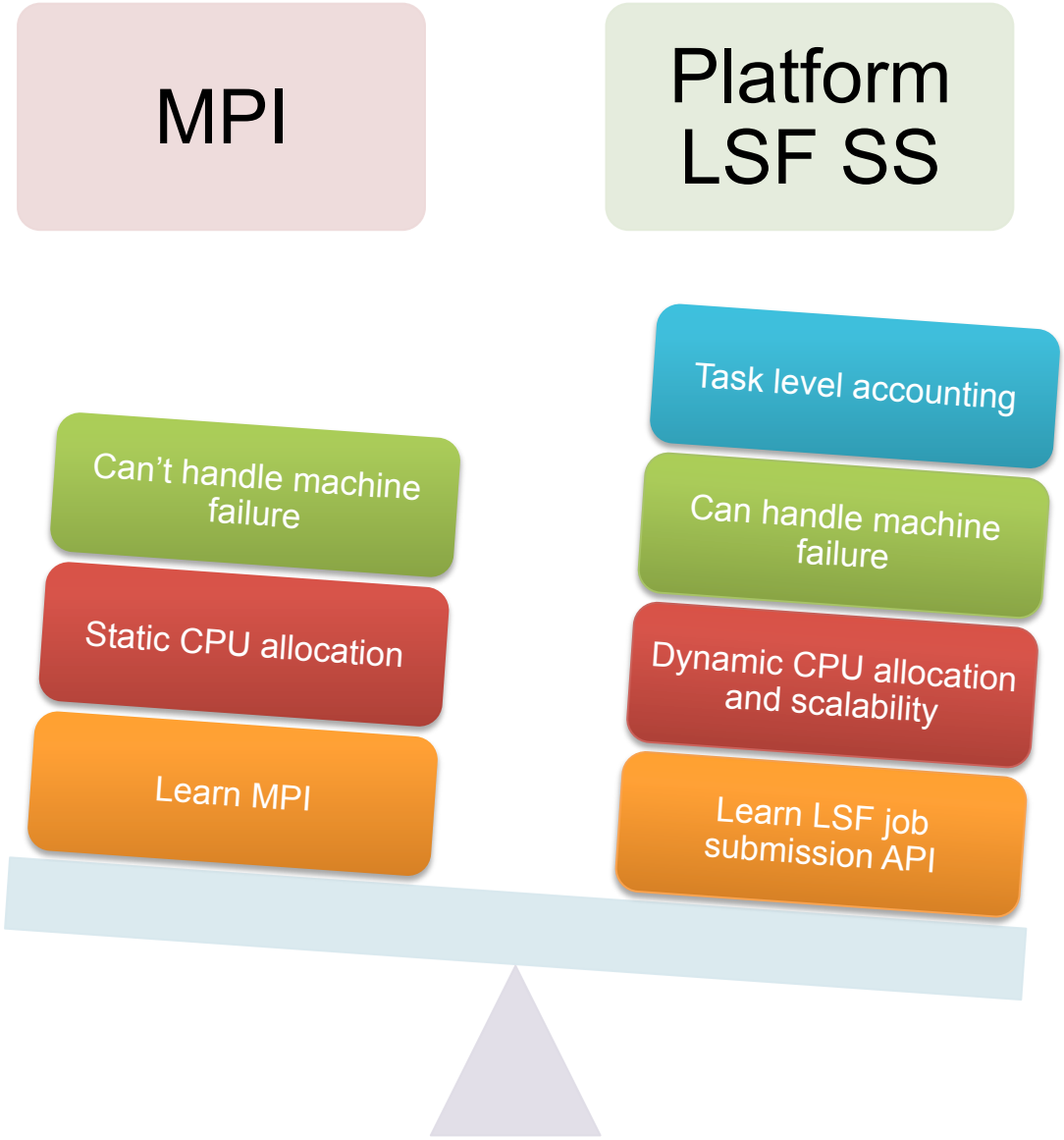
 # bsub –n 100 ssched –task infile

•syntax similar to job arrays
•run extremely large numbers of tasks without impacting the LSF scheduler
•support up to 1,000 simultaneous session schedulers



LSF Session Scheduler

Due to lacking of good task manager, many application developers use MPI to handle embarrassingly parallel tasks



World-class Support & Services

24x7 Support across the globe

"Platform has been proactive, involved and very, very friendly in providing support."

Henry Neeman
Director, Oklahoma University Supercomputing Centre

"Platform's standard of support has been excellent."

Tim Cutts
Platform LSF Administrator
Sanger Institute

Summary

- Platform LSF has extensive support for Multicore
- Platform PCM is now available on the CX1
- Platform LSF session scheduler should be used to efficiently manage high volumes of short jobs
- If you have a workload management problem, we've got a solution!

Platform™

Powering High Performance

www.platform.com

info@platform.com

1-877-528-3676 (1-87-PLATFORM)