SPECS:

Poster Header
40" x 5" TRIM
3 mil lam
1/8" black sintra

COLORS:

Print CMYK

QUANTITY:

1

# STORM Scalable TOol for Resource Management

SPECS:

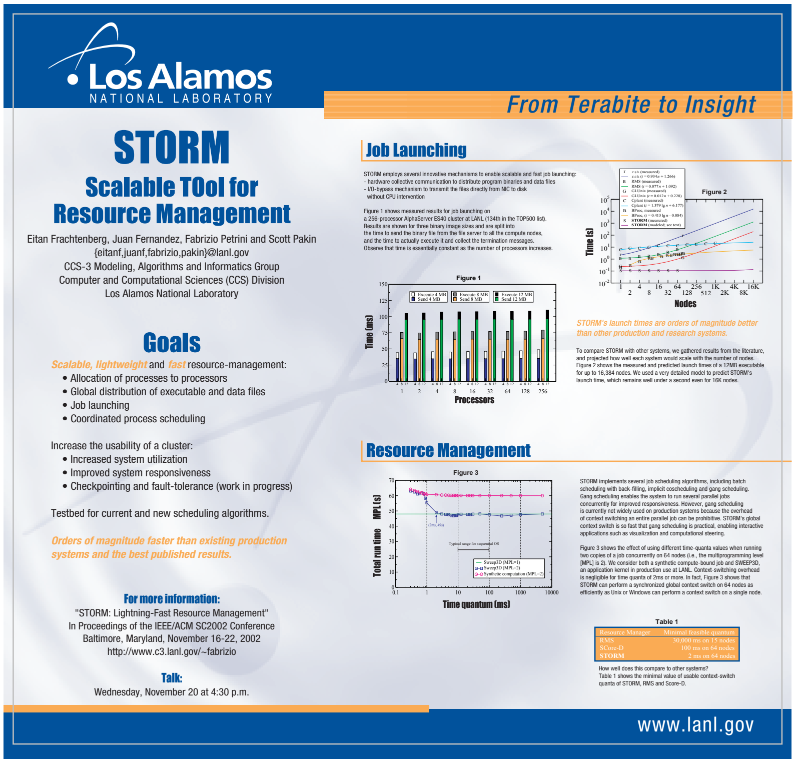
Poster
40" x 38" TRIM
3 mil lam
1/8" black sintra

COLORS:

Print CMYK

QUANTITY:

1



Los Alamos
NATIONAL LABORATORY

*From Terabite to Insight*

# STORM

## Scalable TOol for Resource Management

Eitan Frachtenberg, Juan Fernandez, Fabrizio Petrini and Scott Pakin
{eitanf,juanf,fabrizio,pakin}@lanl.gov
CCS-3 Modeling, Algorithms and Informatics Group
Computer and Computational Sciences (CCS) Division
Los Alamos National Laboratory

## Goals

*Scalable, lightweight* and *fast* resource-management:

- Allocation of processes to processors
- Global distribution of executable and data files
- Job launching
- Coordinated process scheduling

Increase the usability of a cluster:

- Increased system utilization
- Improved system responsiveness
- Checkpointing and fault-tolerance (work in progress)

Testbed for current and new scheduling algorithms.

***Orders of magnitude faster than existing production systems and the best published results.***

**For more information:**

"STORM: Lightning-Fast Resource Management"
In Proceedings of the IEEE/ACM SC2002 Conference
Baltimore, Maryland, November 16-22, 2002
http://www.c3.lanl.gov/~fabrizio

**Talk:**

Wednesday, November 20 at 4:30 p.m.

## Job Launching

STORM employs several innovative mechanisms to enable scalable and fast job launching:
- hardware collective communication to distribute program binaries and data files
- I/O-bypass mechanism to transmit the files directly from NIC to disk without CPU intervention

Figure 1 shows measured results for job launching on a 256-processor AlphaServer ES40 cluster at LANL (134th in the TOP500 list). Results are shown for three binary image sizes and are split into the time to send the binary file from the file server to all the compute nodes, and the time to actually execute it and collect the termination messages. Observe that time is essentially constant as the number of processors increases.

Figure 1

Figure 2

***STORM's launch times are orders of magnitude better than other production and research systems.***

To compare STORM with other systems, we gathered results from the literature, and projected how well each system would scale with the number of nodes. Figure 2 shows the measured and predicted launch times of a 12MB executable for up to 16,384 nodes. We used a very detailed model to predict STORM's launch time, which remains well under a second even for 16K nodes.

## Resource Management

Figure 3

STORM implements several job scheduling algorithms, including batch scheduling with back-filling, implicit coscheduling and gang scheduling. Gang scheduling enables the system to run several parallel jobs concurrently for improved responsiveness. However, gang scheduling is currently not widely used on production systems because the overhead of context switching an entire parallel job can be prohibitive. STORM's global context switch is so fast that gang scheduling is practical, enabling interactive applications such as visualization and computational steering.

Figure 3 shows the effect of using different time-quanta values when running two copies of a job concurrently on 64 nodes (i.e., the multiprogramming level [MPL] is 2). We consider both a synthetic compute-bound job and SWEEP3D, an application kernel in production use at LANL. Context-switching overhead is negligible for time quanta of 2ms or more. In fact, Figure 3 shows that STORM can perform a synchronized global context switch on 64 nodes as efficiently as Unix or Windows can perform a context switch on a single node.

Table 1

| Resource Manager | Minimal feasible quantum |
|---|---|
| RMS | 30,000 ms on 15 nodes |
| SCore-D | 100 ms on 64 nodes |
| **STORM** | 2 ms on 64 nodes |

How well does this compare to other systems?
Table 1 shows the minimal value of usable context-switch quanta of STORM, RMS and Score-D.

www.lanl.gov

LANL

| | | | |
|---|---|---|---|
| CLIENT: | SHOW DATE: | DRAWING TITLE: | DATE: |
| LANL | November 18, 2002 | 40" x 38" adolphy | 11/07/02 |
| PROJECT NAME: | LOCATION: | SCALE: | REVISIONS: |
| Trade Show Booth | ? | 10% of final | 0 of 1 |