Modeling the Impact of Checkpoints on Next-Generation Systems

Cray User Group Technical Conference
May, 2008

SNL	Ron A. Oldfield Rolf Riesen
UTEP	Sarala Arunigiri Patricia Teller Maria Ruiz Varela
IBM	Seetharami Seelam
ORNL	Philip C. Roth

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Fault-Tolerance Challenges for MPP

- MPP Application characteristics
  - Require large fractions of systems (80/40 rule)
  - Long running
  - Resource constrained compute nodes
  - Cannot survive component failure

- Options for fault tolerance
  - Application-directed checkpoints
  - System-directed checkpoints
  - System-directed incremental checkpoints
  - Checkpoint in memory
  - Others: virtualization, redundant computation, …

Application-directed checkpoint to disk dominates!

Sandia National Laboratories

Sandia Fault Tolerance Effort (LDRD)

Questions to answer:

1. **Is checkpoint overhead a real problem for MPPs?**
   - **Account for ~80% of I/O on large systems**
   - **What are current/expected overheads relative to app?**
2. **Can we improve existing approaches?**
3. **Can we contribute a fundamentally different approach?**

This paper/talk addresses the first two questions:

- **Developed analytic model for app-directed chkpt on 3 existing MPPs and one theoretical PetaFlop system**
- **Adapted model to investigate the intermediate nodes as buffers to absorb the "burst" of I/O generated by a checkpoint**

Sandia National Laboratories

Modeling Checkpoint to Disk

- **Goal: Approximate impact of checkpoint to disk on current and future MPP systems**

- **Assume near perfect conditions**
  - **Application uses optimal checkpoint period [Daly]**
  - **Near perfect parallel I/O (at hardware rates)**

Provide a lower bound on the performance impact

(in practice, it will be worse!)

Sandia National Laboratories

The Optimal Checkpoint Interval

- **Daly's equation…**

$$\tau_{opt} = \begin{cases} \sqrt{2\delta M}\left[1+\frac{1}{3}\left(\frac{\delta}{2M}\right)^{1/2}+\frac{1}{9}\left(\frac{\delta}{2M}\right)\right]-\delta & \delta < 2M \\ M & \delta \geq 2M \end{cases}$$

τ_{opt} = Optimal checkpoint interval

δ = Time of the checkpoint operation

M = Mean time to interrupt

- **Not perfect, but it's better than nothing.**

Sandia National Laboratories

Modeling Checkpoints

$$\delta = \alpha_c + \frac{nd}{\min(n\beta_L, \beta_N, \beta_S)}$$

α_c = Start-up overhead of checkpoint

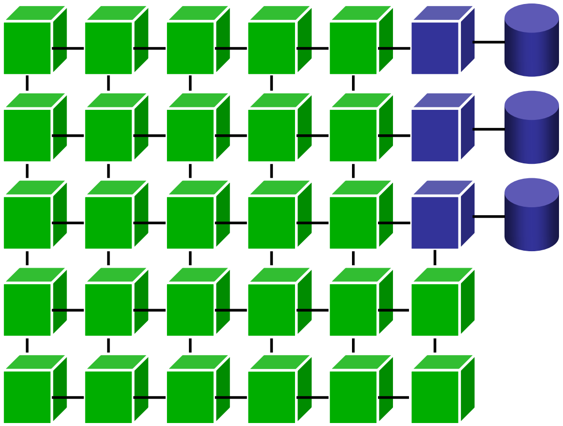
n = Number of compute nodes

d = Data per node dumped to a checkpoint

β_L = Per link bandwidth of the network

β_N = Max network bandwidth to storage

β_S = Aggregate (max) storage bandwidth



Sandia National Laboratories

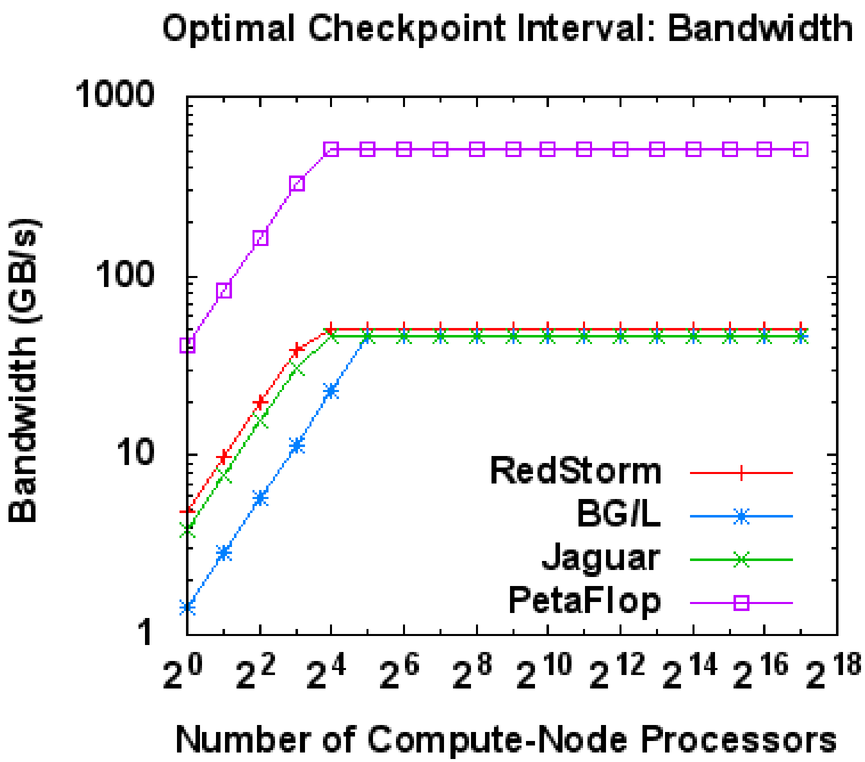
System Parameters

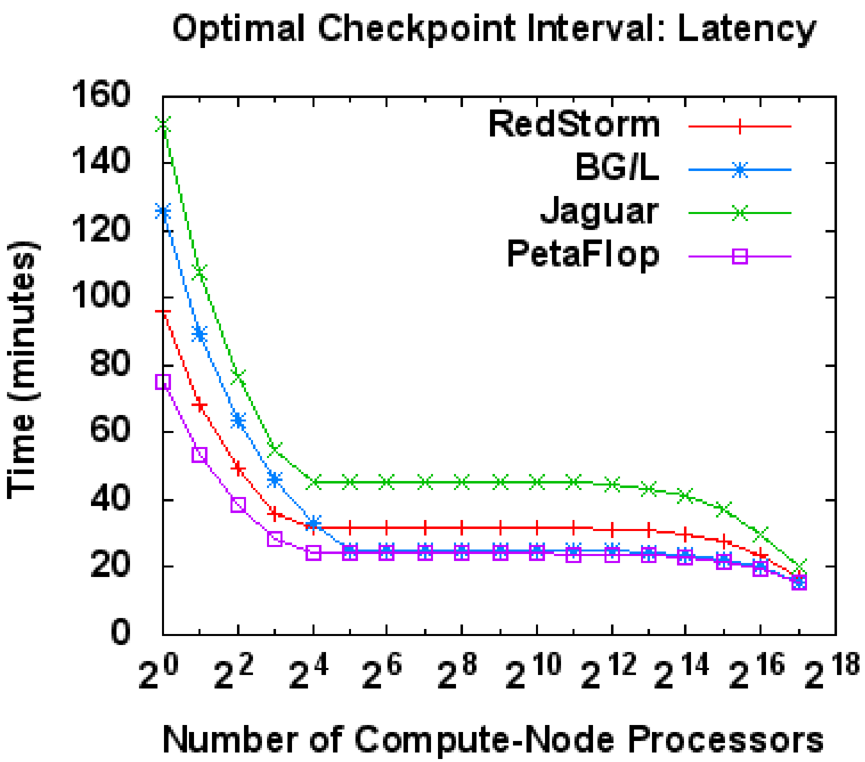
Parameter	Red Storm	BG/L	Jaguar	Petaflop
n (max)	12,960x2	65,536x2	11,590x2	50,000x2
d (max)	1 GB	0.5 GB	2.0 GB	5 GB
*MTTI (dev)**	5 yr	5 yr	5 yr	5 yr
β_S	50 GB/s	45 GB/s	45 GB/s	500 GB/s
β_N	2.3 TB/s	360 GB/s	1.8 TB/s	30 TB/s
β_L	4.8 GB/s	1.4 GB/s	3.8 GB/s	40 GB/s

*** MTTI value comes from a conservative guess based on empirical results (see paper).**

Sandia National Laboratories

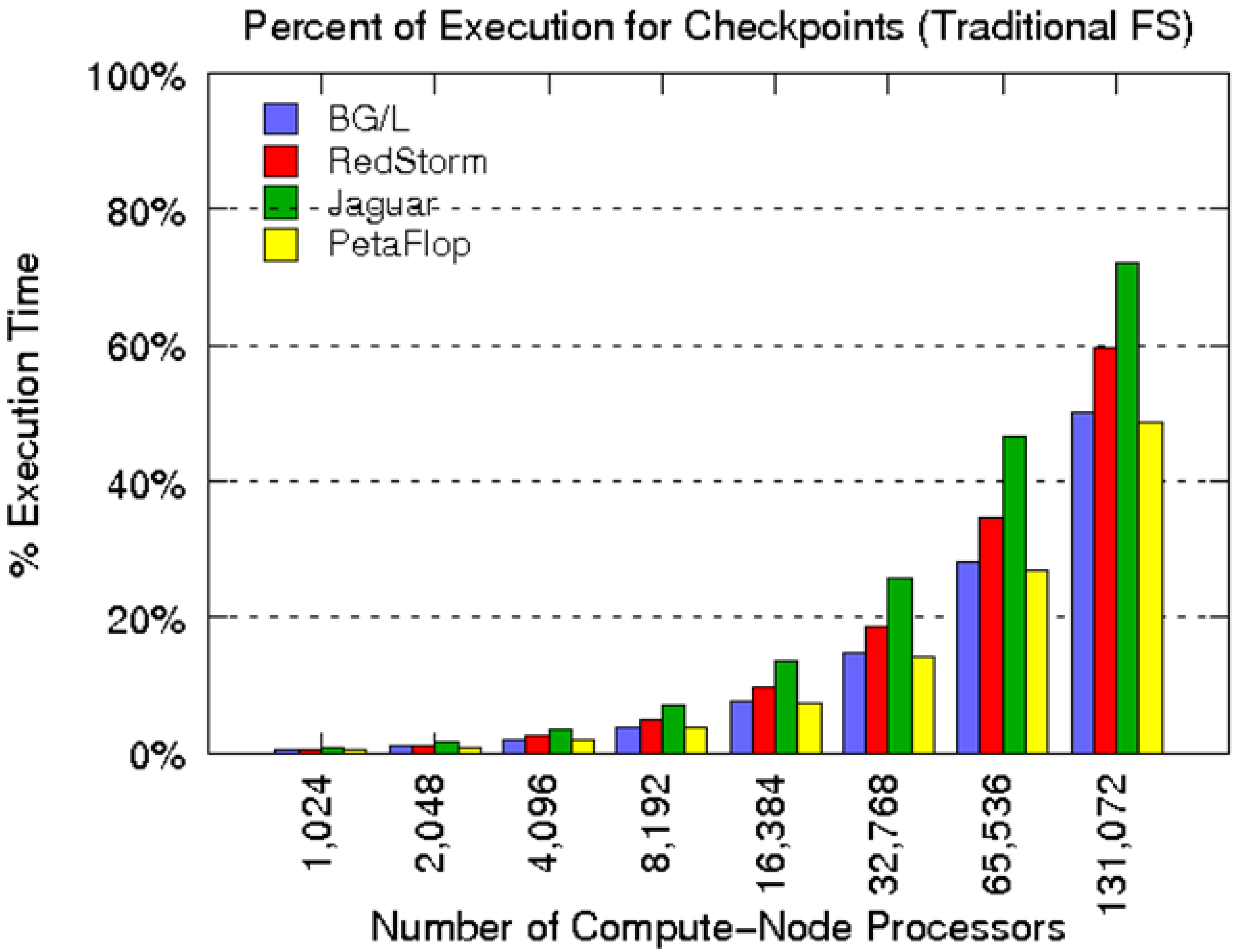
Modeling Results





Sandia National Laboratories

Modeling Results

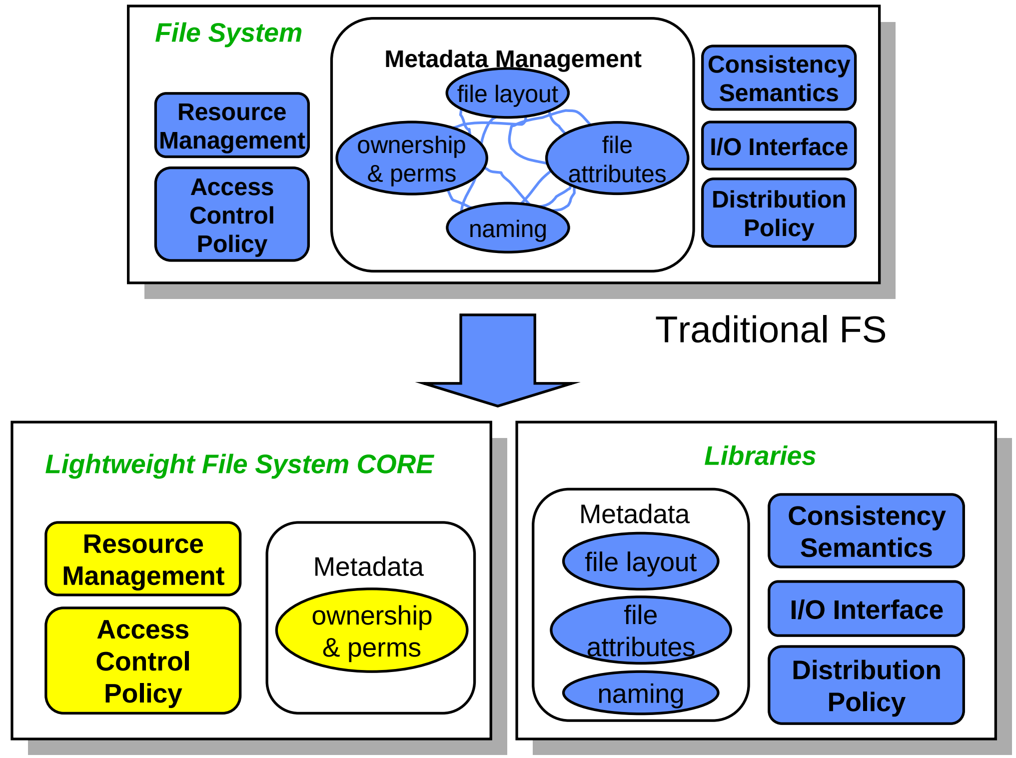


Sandia National Laboratories

Improving I/O Performance of Checkpoints

- Two Proposed Optimizations for MPP Apps
  - The Lightweight File System (LWFS)
  - Use Overlay Networks to absorb I/O bursts

Lightweight File Systems Project



Project Goals

1. **Reduce complexity of FS**
2. **Improve scalability of I/O**

Value of LWFS

- **Vehicle for I/O research**
- **Framework for production FS**
- **Reliable (small code base)**

Cluster'06 paper provides details

LWFS-core Provides

Direct Access to Storage

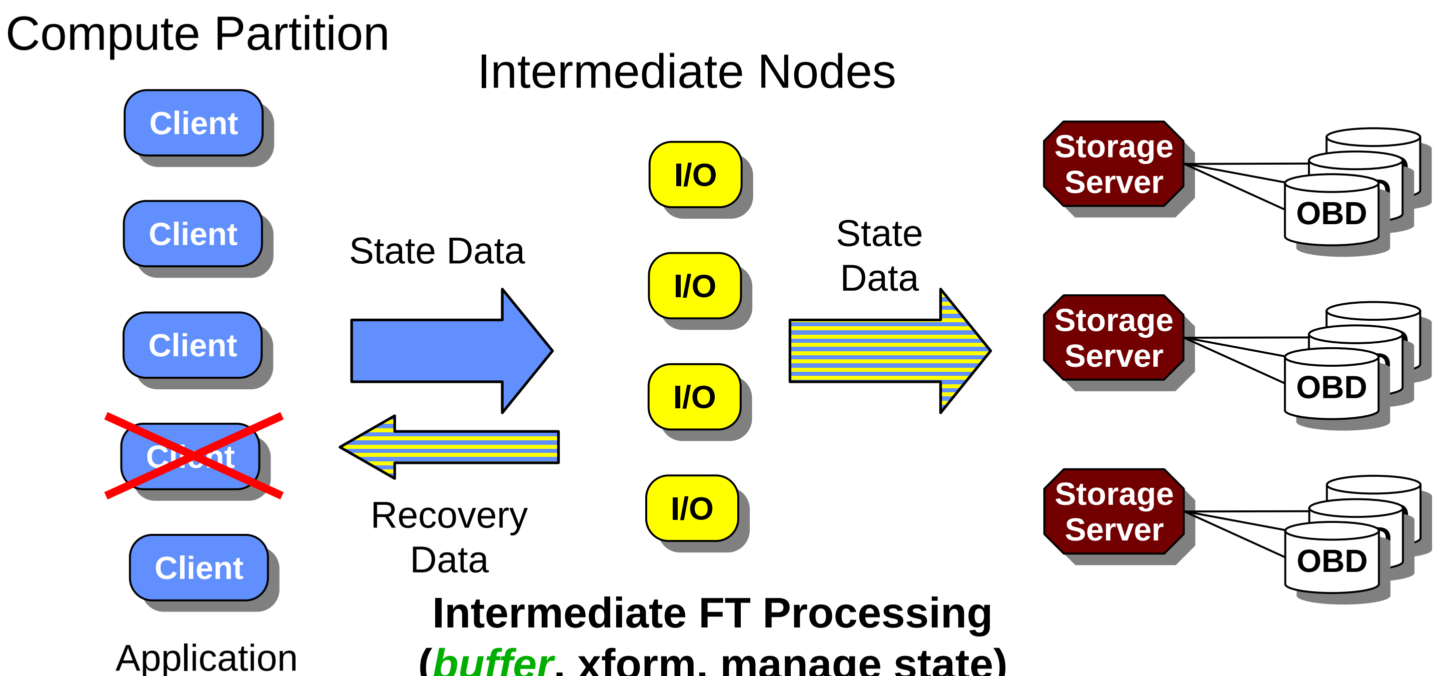
Scalable Security Model

Efficient Data Movement

Libraries Provide

Everything else

LWFS + Overlay Networks



Intermediate FT Processing
(*buffer*, xform, manage state)

Benefits: LWFS + Overlay Network

- Near physical access to storage
- Overlap compute, comm, disk I/O
- Format/permute/partition data for storage
- Manage state for partial application restart

Revisiting the Model for Checkpoints

Bounded by Storage System

Bounded by Network

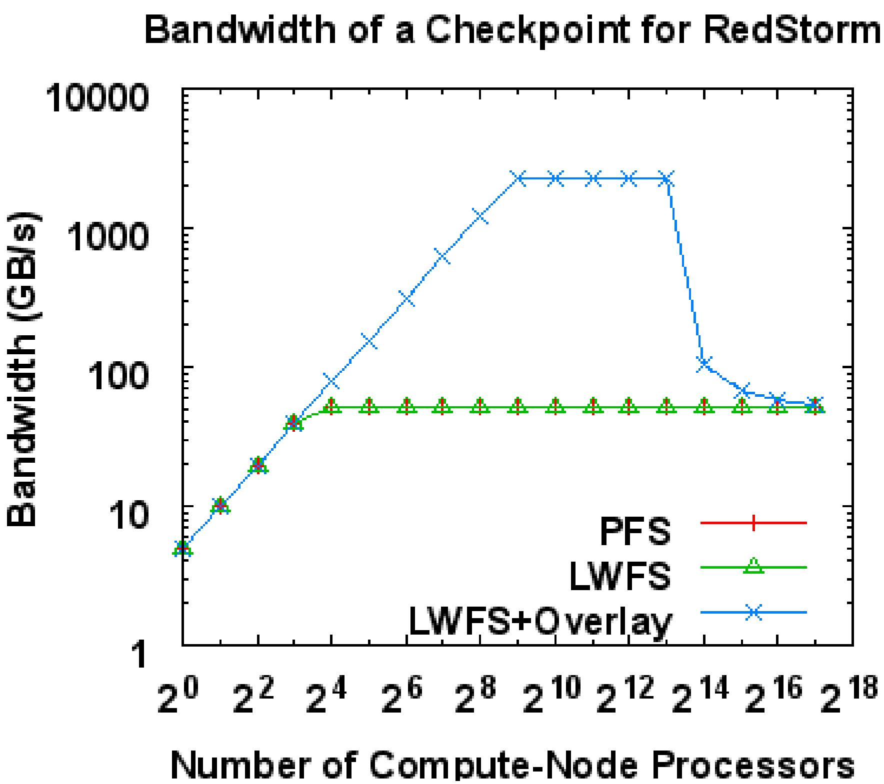
$$\delta = \alpha_c + \begin{cases} \dfrac{nd}{\min(n\beta_L, \beta_N)} & dn \le k \\ \dfrac{k}{\min(n\beta_L, \beta_N)} + \dfrac{nd-k}{\beta_S} & dn > k \end{cases}$$

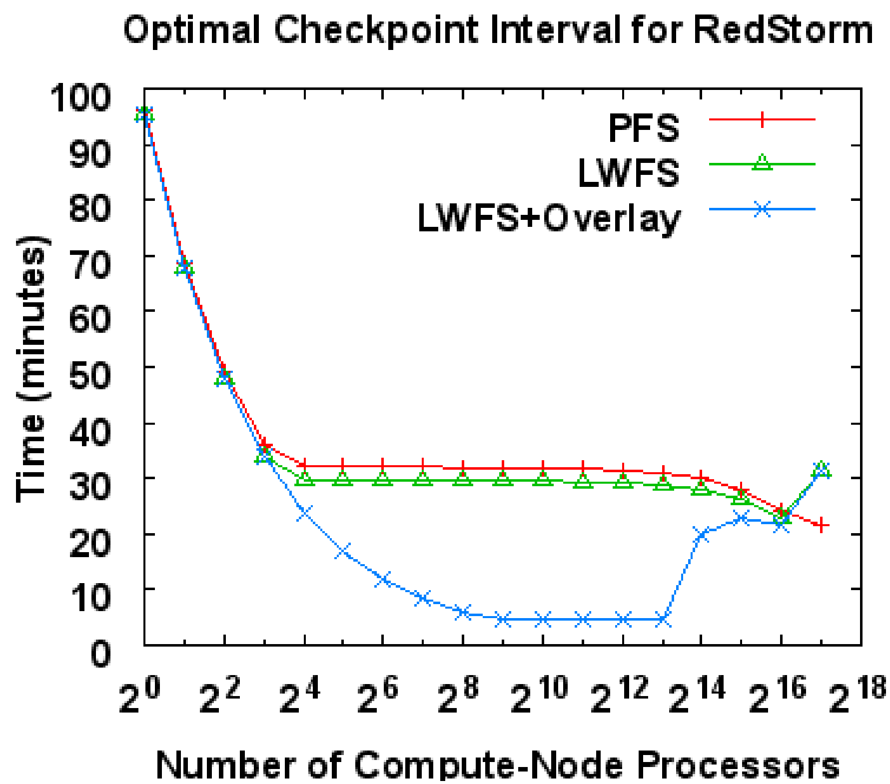
$$k = \mu + \mu\left(\frac{\beta_S}{\min(n\beta_L, \beta_N)}\right) + \cdots = \mu\left(\frac{1}{1 - \dfrac{\beta_S}{\min(n\beta_L, \beta_N)}}\right)$$

$\mu =$ Aggregate memory of intermediate nodes

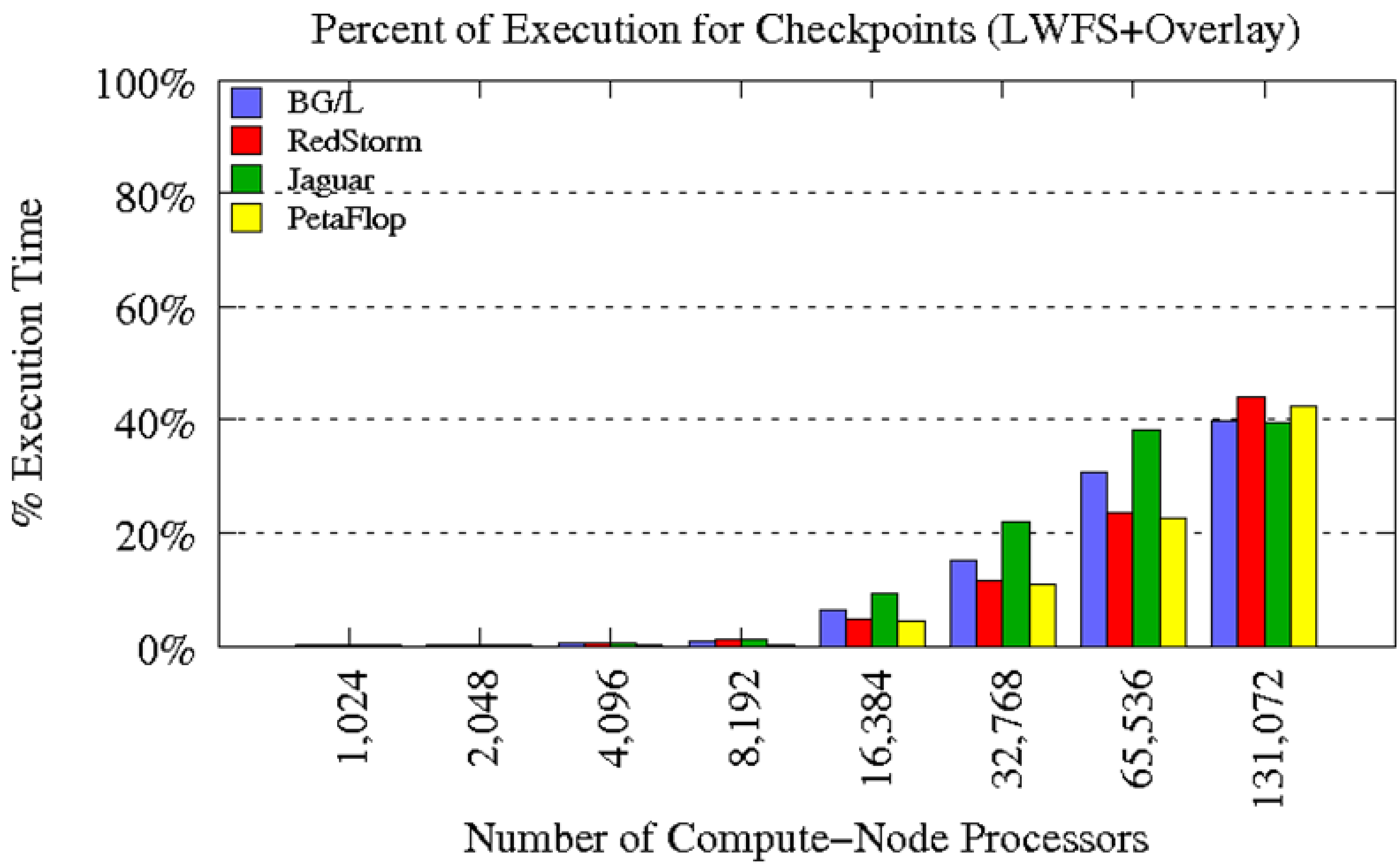
$k =$ Amount of data that can be transferred at network rates

RedStorm Results: PFS, LWFS, and Overlay



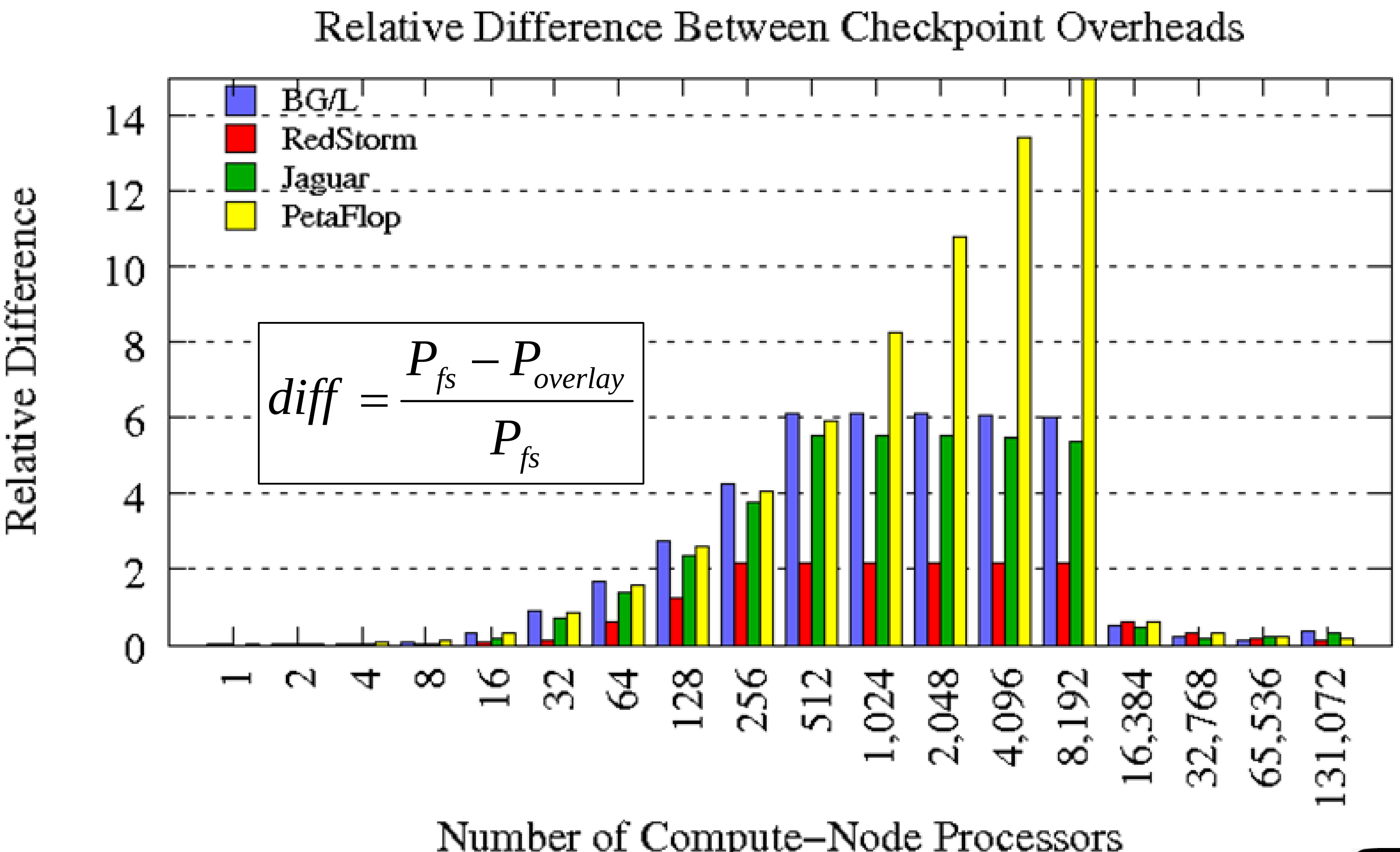


Modeling Results



Sandia National Laboratories

Relative Improvement as a Percentage of Execution Time



Sandia National Laboratories

Summary

- Conclusions from modeling effort
  - Checkpoint to disk is still below "pain threshold"
  - Next-generation systems cause more pain
  - LWFS + Overlays provide some relief
  - "Smart" intermediate nodes could be a cure

- Lots of work to do…
  - Validation of models
  - API's and integration for overlay networks
  - Systems software to support state recovery
  - Algorithms to support state recovery
  - Investigate alternatives to periodic checkpoints
    - Incorporate system info to decide how/when to chkpt (FastOS proposal)

Sandia National Laboratories

Modeling the Impact of Checkpoints on Next-Generation Systems

Cray User Group Technical Conference
May, 2008

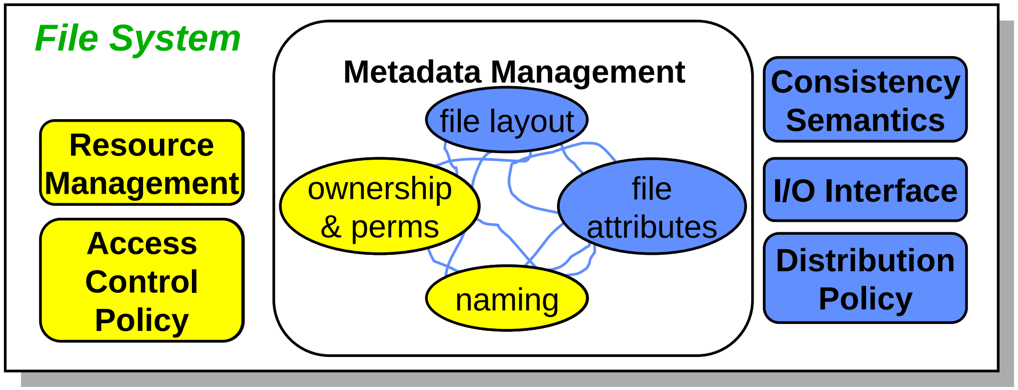
SNL	**Ron A. Oldfield** **Rolf Riesen**
UTEP	**Sarala Arunigiri** **Patricia Teller** **Maria Ruiz Varela**
IBM	**Seetharami Seelam**
ORNL	**Philip C. Roth**

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Extra Slides

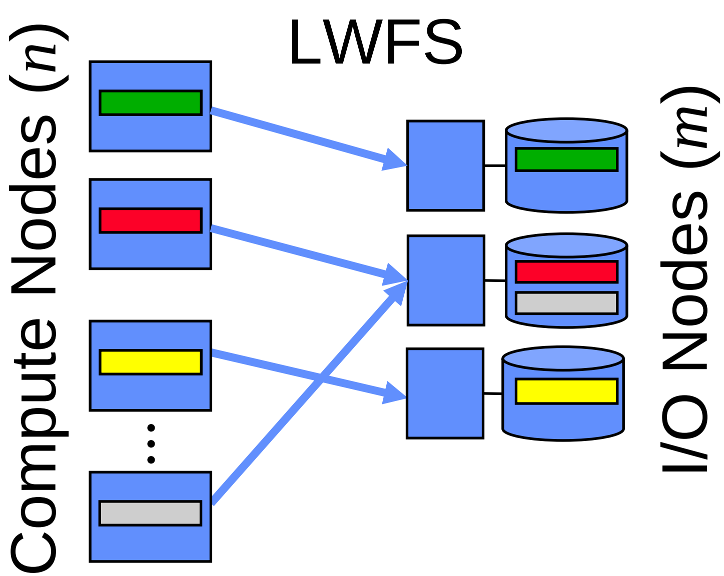
- Advantages of LWFS for Checkpoints
- Additional Results

Checkpoints: Traditional PFS vs. LWFS



Required Operations

	PFS-1 n files nm objs	PFS-2 1 file m objs	LWFS 1 file n objs
create	$n(1+m)$	$m+1$	$n+1$
write	$O(nm)$	$O(nm)$	n



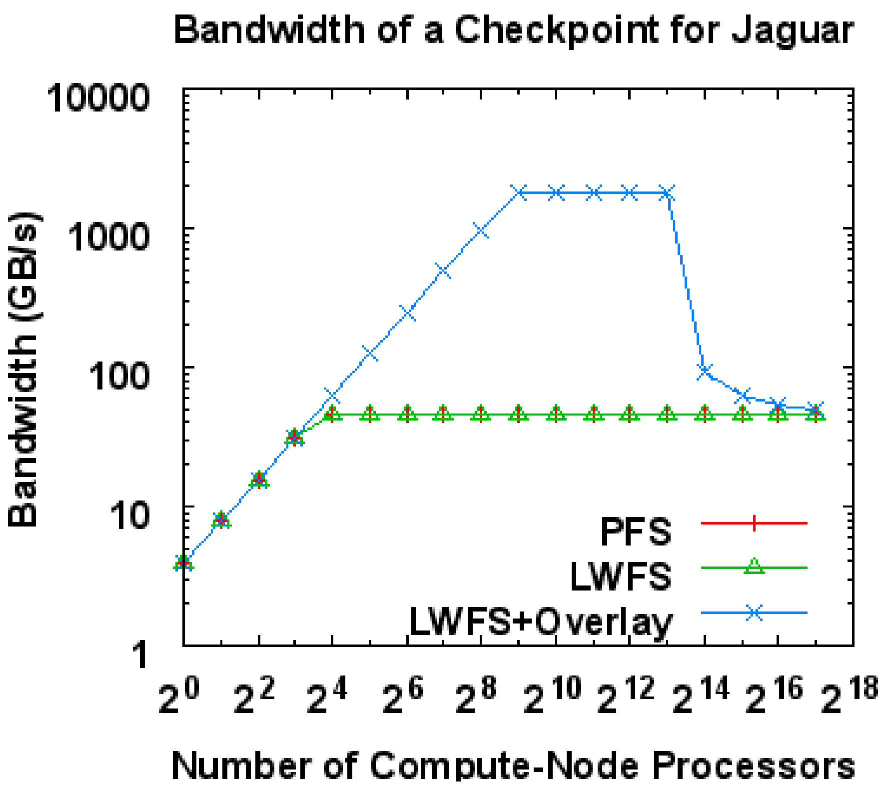
Pseudocode for LWFS

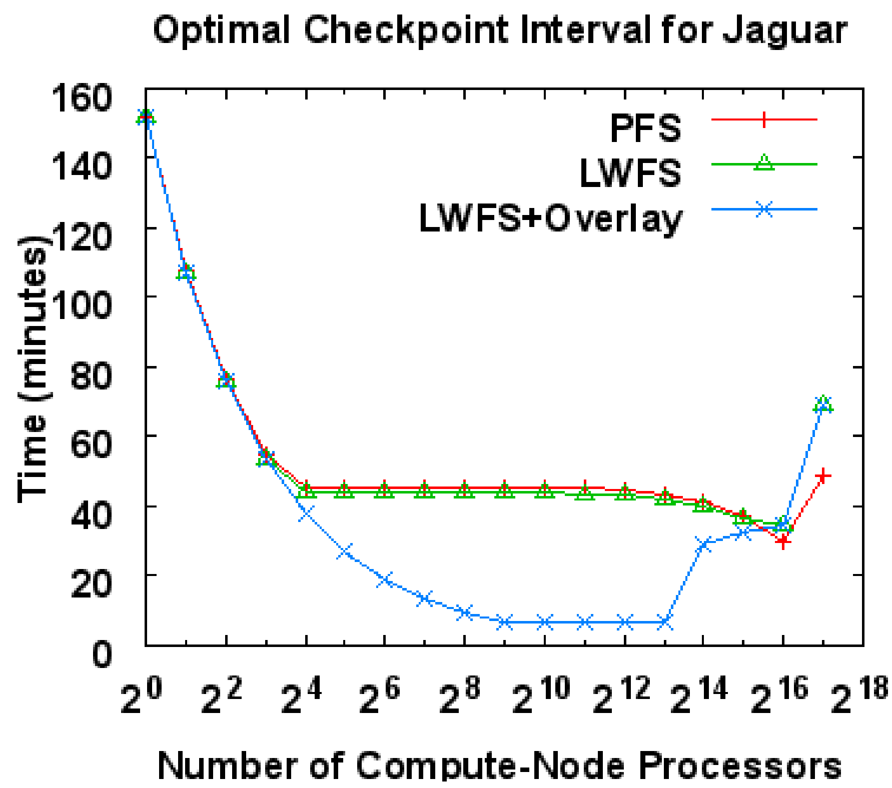
Each Processor (in parallel)
- **Allocate object (blob of bytes)**
- **Dump state**

One processor
- **Allocate object for medata**
- **Gather metadata (obj refs, info about data)**
- **Create name in naming service**
- **Associate MD obj with name**

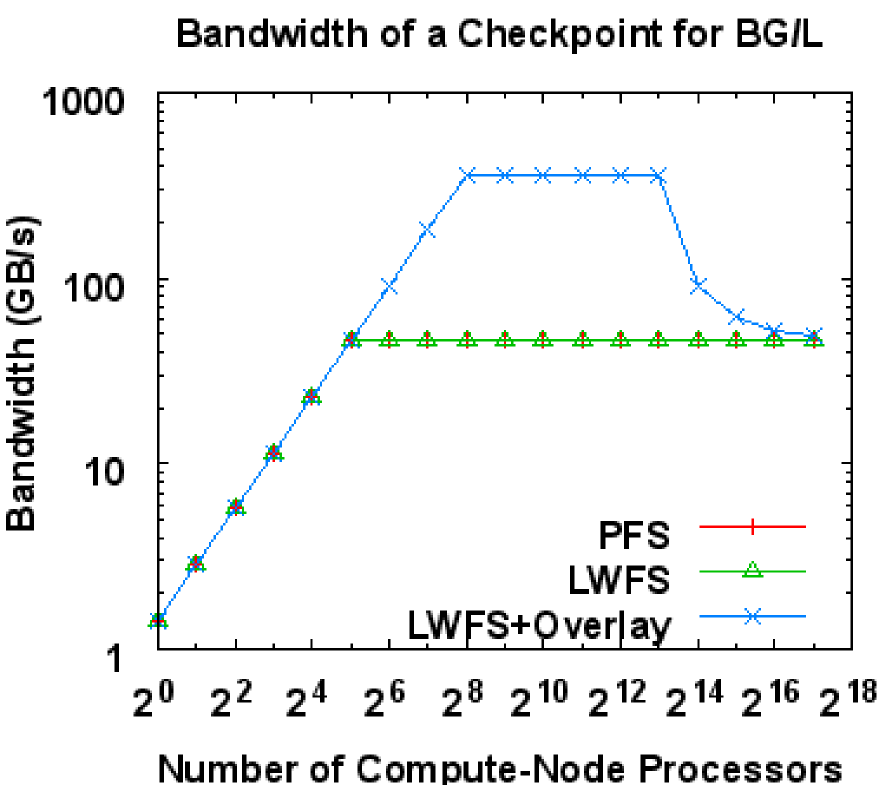
Sandia National Laboratories

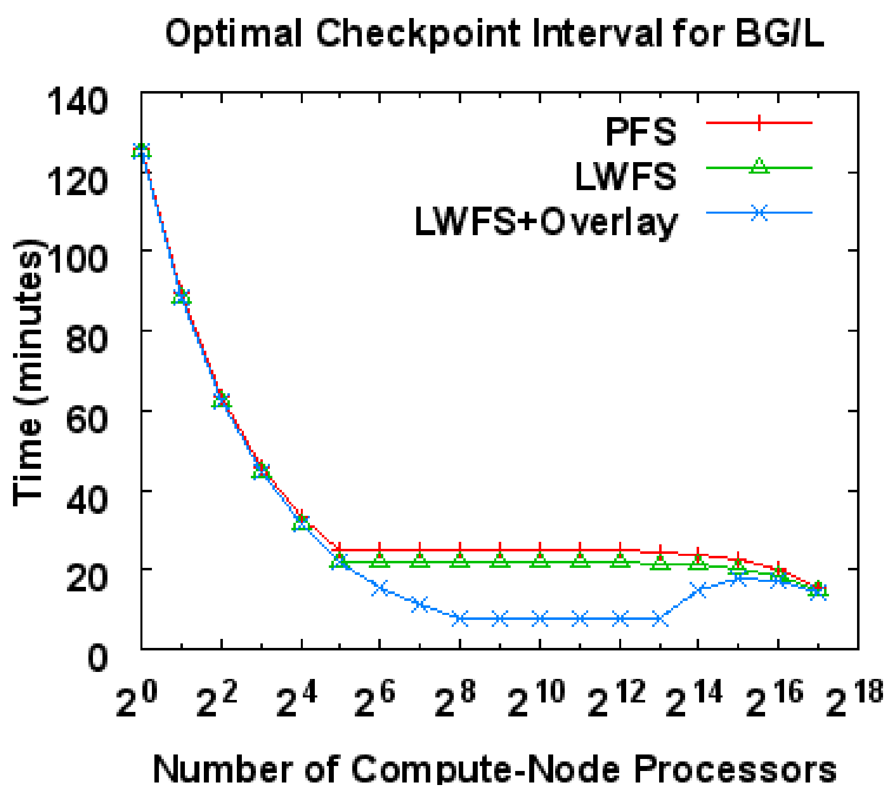
Jaguar Results: PFS, LWFS, and Overlay





BG/L Results: PFS, LWFS, and Overlay





Other results are similar (see extra slides)

Sandia National Laboratories

Petaflop Results: PFS, LWFS, and Overlay