Analysis of the Tradeoffs between Energy and Run Time for Multilevel Checkpointing

Prasanna Balaprakash, **Leonardo A. Bautista Gomez**, Slim Bouguerra, Stefan M. Wild, Franck Cappello , and Paul D. Hovland

ANL

PMBS workshop @ SC'14

Context: The Need For Speed

Systems	2009	2011	2015	2018
System Peak Flops/s	2 Peta	20 Peta	100-200 Peta	1 Exa
System Memory	0.3 PB	1 PB	5 PB	10 PB
Node Performance	125 GF	200 GF	400 GF	1-10 TF
Node Memory BW	25 GB/s	40 GB/s	100 GB/s	200-400 GB/s
Node Concurrency	12	32	O(100)	O(1000)
Interconnect BW	1.5 GB/s	10 GB/s	25 GB/s	50 GB/s
System Size (Nodes)	18,700	100,000	500,000	O(Million)
Total Concurrency	225,000	3 Million	50 Million	O(Billion)
Storage	15 PB	30 PB	150 PB	300 PB
I/O	0.2 TB/s	2 TB/s	10 TB/s	20 TB/s
MTTI	Days	Days	Days	O(1Day)
Power	6 MW	~10 MW	~10 MW	~20 MW

Figure : From http://www.scidacreview.org/

Motivation: Failures

- Sequoia MTBF $\approx$ 1 day.
- Blue Waters 2 nodes failure per day.
- Titan MTBF $<$ 1 day.

Motivation: Failures

- Sequoia MTBF $\approx$ 1 day.
- Blue Waters 2 nodes failure per day.
- Titan MTBF $<$ 1 day.
- $\approx$ 20 % of the computation is wasted in recovery and re-execution (Implies energy waste)

Motivation: Failures

- Sequoia MTBF $\approx$ 1 day.
- Blue Waters 2 nodes failure per day.
- Titan MTBF $<$ 1 day.
- $\approx$ 20 % of the computation is wasted in recovery and re-execution (Implies energy waste)

Exascale:

- The number of components for both memory and processors will increase by a factor of 100.
- Shrinking the circuit sizes and running at lower voltages, increases the SDC probability.

Motivation: Failures

- Sequoia MTBF $\approx$ 1 day.
- Blue Waters 2 nodes failure per day.
- Titan MTBF $<$ 1 day.
- $\approx$ 20 % of the computation is wasted in recovery and re-execution (Implies energy waste)

Exascale:

- The number of components for both memory and processors will increase by a factor of 100.
- Shrinking the circuit sizes and running at lower voltages, increases the SDC probability.

In exascale failures will occur at higher frequency, optimistic MTBF is couple of hours.

Motivation: Energy

- The power draw of the interconnect on Blue Gene/Q appears to be independent of load.
- CPU varies only by some 20%
- Power draw under different loads is DRAM change by a factor of 2 or more.

Exascale: http://www.scidacreview.org/1001/html/hardware.html

Data movement and IO will consume more than 70% of the total system power (most of the 20 MW will go just to power the 10 PB of total system memory.)

Flops/Watt VS Communication/Watts

Avoid checkpointing and data movement do more re-computations.

VS

Avoid re-computations via checkpointing more often.

Related Work

- ECOTFIT, Diouri et al
  - Blocking checkpointing
  - Message logging
  - Conclusion no big tradeoff observed.
- Meneses et al
  - Parallel recovery vs global recovery
  - Used RALP API (No communication or IO are covered)
  - Parallel is better since it reduces the overall time
- Aupy et al
  - Blocking vs no-blocking single level.
  - No experiment.

Related Work

- ECOTFIT, Diouri et al
  - Blocking checkpointing
  - Message logging
  - Conclusion no big tradeoff observed.
- Meneses et al
  - Parallel recovery vs global recovery
  - Used RALP API (No communication or IO are covered)
  - Parallel is better since it reduces the overall time
- Aupy et al
  - Blocking vs no-blocking single level.
  - No experiment.

The missing episode

What about multilevel checkpointing ?

1. Context and motivations
2. Problem formulation and notations
   - Multilevel checkpointing
   - Energy model
   - Multiobjective optimization
3. Simulation and experimentations
   - Experimentations
   - Tradeoff analysis
4. Conclusion and future work

Multilevel Checkpointing

- Multiple levels of storage (DRAM, NVM, PFS).
- Coupled with data replication and erasure codes.
- Low levels offer high performance and partial reliability.
- High levels offer high reliability but impose large overhead.
- Different ckpt. levels have different frequencies.
- After a failure the application restart from the lowest available level.
- If unable to recover, try next level of checkpoint (Further in the past).

Wasted time model

- L levels of checkpoint (4 with FTI)
- Checkpoint strategy: τ_i, for $i = 1 \cdots L$
- Checkpoint cost: c_i for level i
- r_i time for a restart from level i
- d_i downtime after a failure affecting level i.
- μ_i rate of failures affecting level i.

Wasted energy model

- $\mathcal{P}_i^c$ power for a level i checkpoint Watts.
- $\mathcal{P}_i^r$ power for a restart from level i Watts.
- $\mathcal{P}^a$ power for a failure-free computation without checkpointing Watts.
- μ_i rate for failure affecting level i.

Problem solving

Checkpoint time

$$\mathfrak{W}^{ch} = \sum_{i=1}^{L} \left(\frac{c_i}{\tau_i} + \mu_i \tau_i \sum_{j=1}^{i-1} \frac{c_j}{2\tau_j} \right)$$

Problem solving

Checkpoint time

$$\mathfrak{W}^{ch} = \sum_{i=1}^{L} \left(\frac{c_i}{\tau_i} + \mu_i \tau_i \sum_{j=1}^{i-1} \frac{c_j}{2\tau_j} \right)$$

Rework time

$$\mathfrak{W}^{rew} = \sum_{i=1}^{L} \frac{\mu_i \tau_i}{2}$$

Problem solving

Checkpoint time

$$\mathfrak{W}^{ch} = \sum_{i=1}^{L} \left(\frac{c_i}{\tau\ i} + \mu_i \tau_i \sum_{j=1}^{i-1} \frac{c_j}{2\tau_j} \right)$$

Rework time

$$\mathfrak{W}^{rew} = \sum_{i=1}^{L} \frac{\mu_i \tau_i}{2}$$

Downtime and restart time

$$\mathfrak{W}^{down} = \sum_{i=1}^{L} \mu_i (r_i + d_i)$$

Problem solving

Checkpoint wasted energy

$$\mathcal{E}^{ch} = \sum_{i=1}^{L} \mathcal{P}_i^c \frac{c_i}{\tau_i} + \mu_i \tau_i \sum_{j=1}^{i-1} \frac{\mathcal{P}_j^c c_j}{2\tau_j}$$

Problem solving

Checkpoint wasted energy

$$\mathcal{E}^{ch} = \sum_{i=1}^{L} \mathcal{P}_i^c \frac{c_i}{\tau_i} + \mu_i \tau_i \sum_{j=1}^{i-1} \frac{\mathcal{P}_j^c c_j}{2\tau_j}$$

Rework wasted energy

$$\mathcal{E}^{rew} = \sum_{i=1}^{L} \mathcal{P}^a \frac{\mu_i \tau_i}{2}$$

Problem solving

Checkpoint wasted energy

$$\mathcal{E}^{ch} = \sum_{i=1}^{L} \mathcal{P}_i^c \frac{c_i}{\tau_i} + \mu_i \tau_i \sum_{j=1}^{i-1} \frac{\mathcal{P}_j^c c_j}{2\tau_j}$$

Rework wasted energy

$$\mathcal{E}^{rew} = \sum_{i=1}^{L} \mathcal{P}^a \frac{\mu_i \tau_i}{2}$$

Downtime and restart wasted energy

$$\mathcal{E}^{down} = \sum_{i=1}^{L} \mathcal{P}_i^r \mu_i (r_i + d_i)$$

Total wasted time

$$\mathbb{W} = \sum_{i=1}^{L} \left(\frac{c_i}{\tau_i} + \frac{\mu_i \tau_i}{2} \left(1 + \sum_{j=1}^{i-1} \frac{c_j}{2\tau_j} \right) + \mu_i \left(r_i + d_i \right) \right) \quad (1)$$

Total wasted time

$$\mathbb{W} = \sum_{i=1}^{L} \left(\frac{c_i}{\tau_i} + \frac{\mu_i \tau_i}{2} \left(1 + \sum_{j=1}^{i-1} \frac{c_j}{2\tau_j} \right) + \mu_i \left(r_i + d_i \right) \right) \quad (1)$$

Total wasted energy

$$\mathbb{E} = \sum_{i=1}^{L} \left(\frac{\mathcal{P}_i^c c_i}{\tau_i} + \mu_i \tau_i \left(\frac{\mathcal{P}^a}{2} + \sum_{j=1}^{i-1} \frac{\mathcal{P}_j^c c_j}{2\tau_j} \right) \right) + \sum_{i=1}^{L} \mathcal{P}_i^r \mu_i (r_i + d_i), \quad (2)$$

First derivatives

$$\frac{\partial \mathbb{W}}{\partial \tau_i} = \frac{\mu_i}{2}\left(1 + \sum_{j=1}^{i-1} \frac{c_j}{\tau_j}\right) - \frac{c_i}{\tau_i^2}\left(1 + \sum_{j=i+1}^{L} \frac{\mu_j \tau_j}{2}\right) \tag{3}$$

$$\frac{\partial \mathbb{E}}{\partial \tau_i} = \frac{\mu_i}{2}\left(\mathcal{P}^a + \sum_{j=1}^{i-1} \frac{\mathcal{P}_j^c c_j}{\tau_j}\right) - \frac{\mathcal{P}_i^c c_i}{\tau_i^2}\left(1 + \sum_{j=i+1}^{L} \frac{\mu_j \tau_j}{2}\right) \tag{4}$$

First derivatives

$$\frac{\partial \mathbb{W}}{\partial \tau_i} = \frac{\mu_i}{2}\left(1 + \sum_{j=1}^{i-1} \frac{c_j}{\tau_j}\right) - \frac{c_i}{\tau_i^2}\left(1 + \sum_{j=i+1}^{L} \frac{\mu_j \tau_j}{2}\right) \tag{3}$$

$$\frac{\partial \mathbb{E}}{\partial \tau_i} = \frac{\mu_i}{2}\left(\mathcal{P}^a + \sum_{j=1}^{i-1} \frac{\mathcal{P}_j^c c_j}{\tau_j}\right) - \frac{\mathcal{P}_i^c c_i}{\tau_i^2}\left(1 + \sum_{j=i+1}^{L} \frac{\mu_j \tau_j}{2}\right) \tag{4}$$

Solutions

$$\tau_i^{\mathbb{W}} = \sqrt{\frac{c_i(2 + \sum_{j=i+1}^{L} \mu_j \tau_j^{\mathbb{W}})}{\mu_i(1 + \sum_{j=1}^{i-1} \frac{c_j}{\tau_j^{\mathbb{W}}})}}$$

$$\tau_i^{\mathbb{E}} = \sqrt{\frac{\rho_i c_i(2 + \sum_{j=i+1}^{L} \mu_j \tau_j^{\mathbb{E}})}{\mu_i(1 + \sum_{j=1}^{i-1} \frac{\rho_j c_j}{\tau_j^{\mathbb{E}}})}}$$

$\rho_i = \mathcal{P}_i^c / \mathcal{P}^a$

Solutions

$$\tau_i^{\mathbb{W}} = \sqrt{\frac{c_i(2 + \sum_{j=i+1}^{L} \mu_j \tau_j^{\mathbb{W}})}{\mu_i(1 + \sum_{j=1}^{i-1} \frac{c_j}{\tau_j^{\mathbb{W}}})}}$$

$$\tau_i^{\mathbb{E}} = \sqrt{\frac{\rho_i c_i(2 + \sum_{j=i+1}^{L} \mu_j \tau_j^{\mathbb{E}})}{\mu_i(1 + \sum_{j=1}^{i-1} \frac{\rho_j c_j}{\tau_j^{\mathbb{E}}})}}$$

$\rho_i = \mathcal{P}_i^c / \mathcal{P}^a$

Solutions

For one single level we have : $\tau^{\mathbb{W}} = \sqrt{2c/\mu}$ and $\tau^{\mathbb{E}} = \tau^{\mathbb{W}} \sqrt{\mathcal{P}^c / \mathcal{P}^a}$
Whenever $\mathcal{P}^c \neq \mathcal{P}^a$, we have that $\tau^{\mathbb{W}} \neq \tau^{\mathbb{E}}$, and hence the two objectives are conflicting.

Pareto front

Definition

τ^i is said to be *Pareto-optimal* if it is not dominated by any other τ^j.

Pareto front

Definition

τ^i is said to be *Pareto-optimal* if it is not dominated by any other τ^j.

Convex combination

If the Pareto front is convex, any point on the front can be obtained by minimizing a linear combination of the objectives.

$$f_\lambda(\tau) = \lambda \mathbb{W}(\tau) + (1 - \lambda)\mathbb{E}(\tau), \text{for } \lambda \in [0, 1].$$

Theorem

The Hessian $\nabla^2_{\tau\tau}\mathbb{W}(\tau)$ is diagonally dominant, and thus $\mathbb{W}$ is a convex function of τ over the domain (same for $\mathbb{E}(\tau)$)

Pareto front

Convex combination

If the Pareto front is convex, any point on the front can be obtained by minimizing a linear combination of the objectives.

$$f_\lambda(\tau) = \lambda \mathbb{W}(\tau) + (1-\lambda)\mathbb{E}(\tau), \text{for } \lambda \in [0,1].$$

Theorem

The Hessian $\nabla^2_{\tau\tau}\mathbb{W}(\tau)$ is diagonally dominant, and thus $\mathbb{W}$ is a convex function of τ over the domain (same for $\mathbb{E}(\tau)$)

$$\tau_i^*(\lambda) = \sqrt{\frac{c_i(\lambda + (1-\lambda)\mathcal{P}_i^c)\left(2 + \sum_{j=i+1}^{L} \mu_j \tau_j^*\right)}{\mu_i\left(\lambda + (1-\lambda)\mathcal{P}^a + \sum_{j=1}^{i-1}(\lambda + (1-\lambda)\mathcal{P}_j^c)\frac{c_j}{\tau_j^*}\right)}}, \tag{5}$$

Pareto front

Theorem

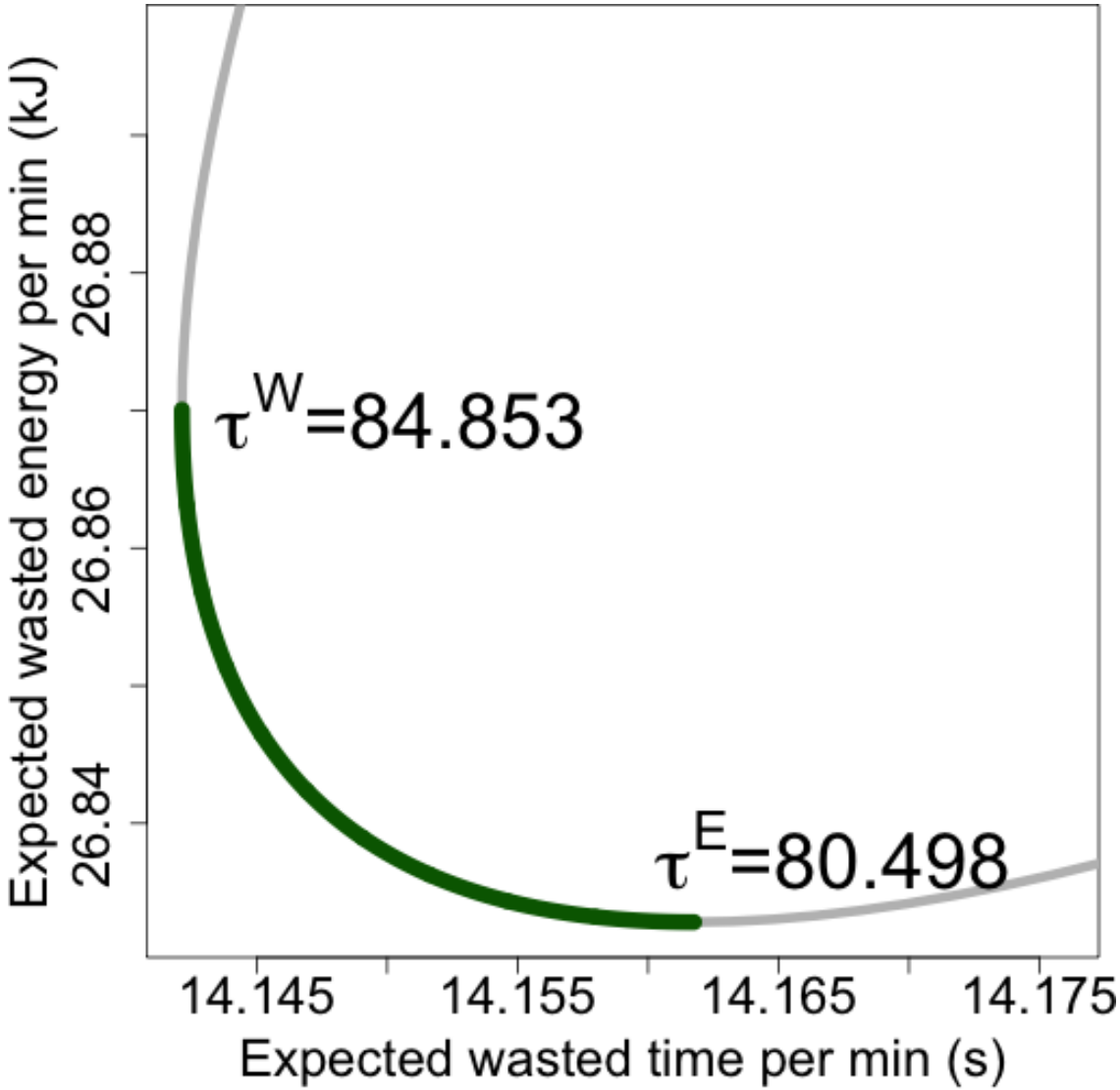
The Hessian $\nabla^2_{\tau\tau}\mathbb{W}(\tau)$ is diagonally dominant, and thus $\mathbb{W}$ is a convex function of τ over the domain (same for $\mathbb{E}(\tau)$)

$$\tau_i^*(\lambda) = \sqrt{\frac{c_i(\lambda + (1-\lambda)\mathcal{P}_i^c)\left(2 + \sum_{j=i+1}^{L} \mu_j \tau_j^*\right)}{\mu_i\left(\lambda + (1-\lambda)\mathcal{P}^a + \sum_{j=1}^{i-1}(\lambda + (1-\lambda)\mathcal{P}_j^c)\frac{c_j}{\tau_j^*}\right)}}, \tag{5}$$

Case one level ($L = 1$)

$$\tau^*(\lambda) = \tau^{\mathbb{W}}\sqrt{\frac{\lambda + (1-\lambda)\mathcal{P}^c}{\lambda + (1-\lambda)\mathcal{P}^a}}. \tag{6}$$

Pareto Front



1. Context and motivations
2. Problem formulation and notations
   - Multilevel checkpointing
   - Energy model
   - Multiobjective optimization
3. Simulation and experimentations
   - Experimentations
   - Tradeoff analysis
4. Conclusion and future work

Platform

- Mira 10 PF IBM Blue Gene/Q (BG/Q)
  - 49,152 nodes organized in 48 racks
  - 16 cores of 1.6 GHz PowerPC A2 and 16 GB of DDR3 memory.
  - 5-D torus network.
- Vesta (developmental platform for Mira)
  - Same as Mira's but with 2,048 nodes

Platform

- Mira 10 PF IBM Blue Gene/Q (BG/Q)
  - 49,152 nodes organized in 48 racks
  - 16 cores of 1.6 GHz PowerPC A2 and 16 GB of DDR3 memory.
  - 5-D torus network.
- Vesta (developmental platform for Mira)
  - Same as Mira's but with 2,048 nodes

MonEQ for power measurement (resolution of 560 ms)

- Chip core
- DRAM
- Network
- Collect power data only at the node card level (every 32 nodes)
- The library can not measure the I/O power consumption !!

Applications

- LAMMPS
  - Production-level molecular dynamics application.
  - Lennard-Jones simulation of 1.3 billion atoms.
  - Checkpoint size per node $\approx$ 200 MB ($\approx$100GB application's footprint)
  - Checkpoints intervals 4, 8, 16 and 32 minutes.

- CORAL
  - Qbox: first-principles molecular dynamics
  - AMG: is a parallel algebraic multigrid solver for linear systems arising from problems on unstructured grids.
  - LULESH: performs hydrodynamics stencil calculations
  - miniFE: is a finite-element code.

LAMMPS: synchronous

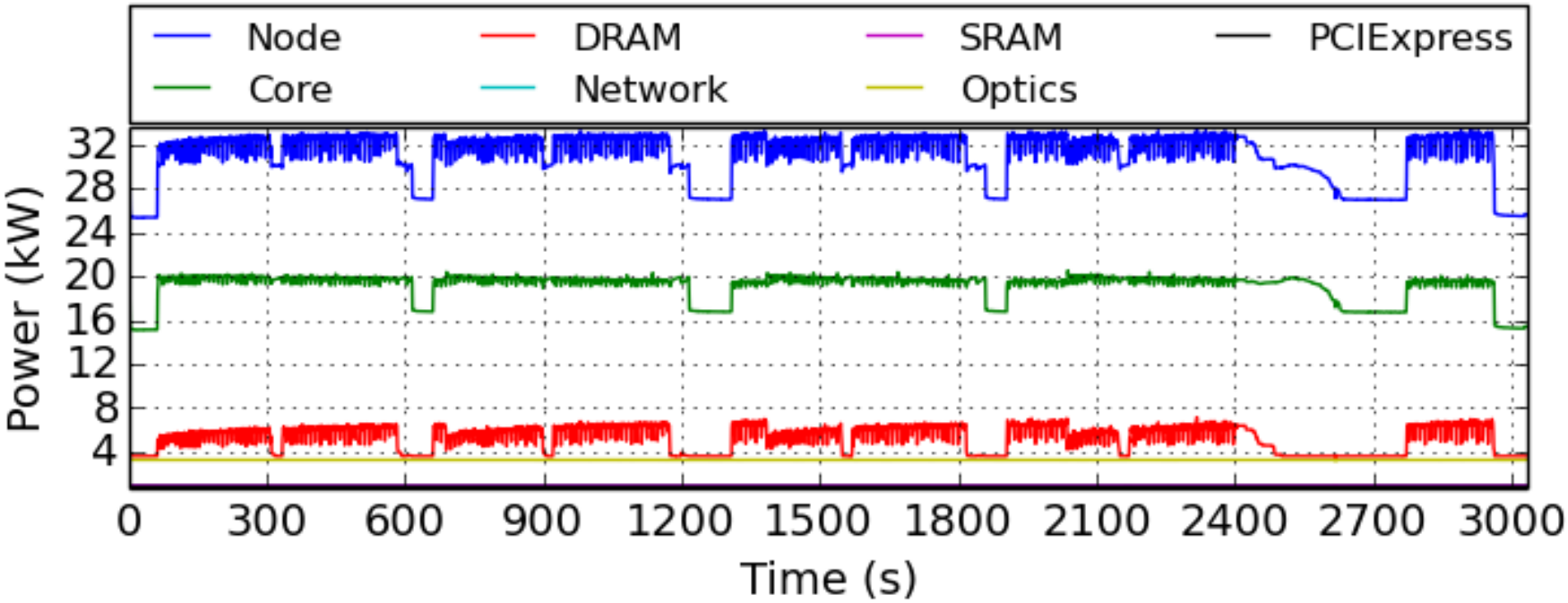


Figure : Synchronous multilevel checkpointing

1.3 billion atoms Lennard-Jones simulation.
512 nodes running 64 MPI ranks per node (32,678 proc.).

LAMMPS: asynchronous

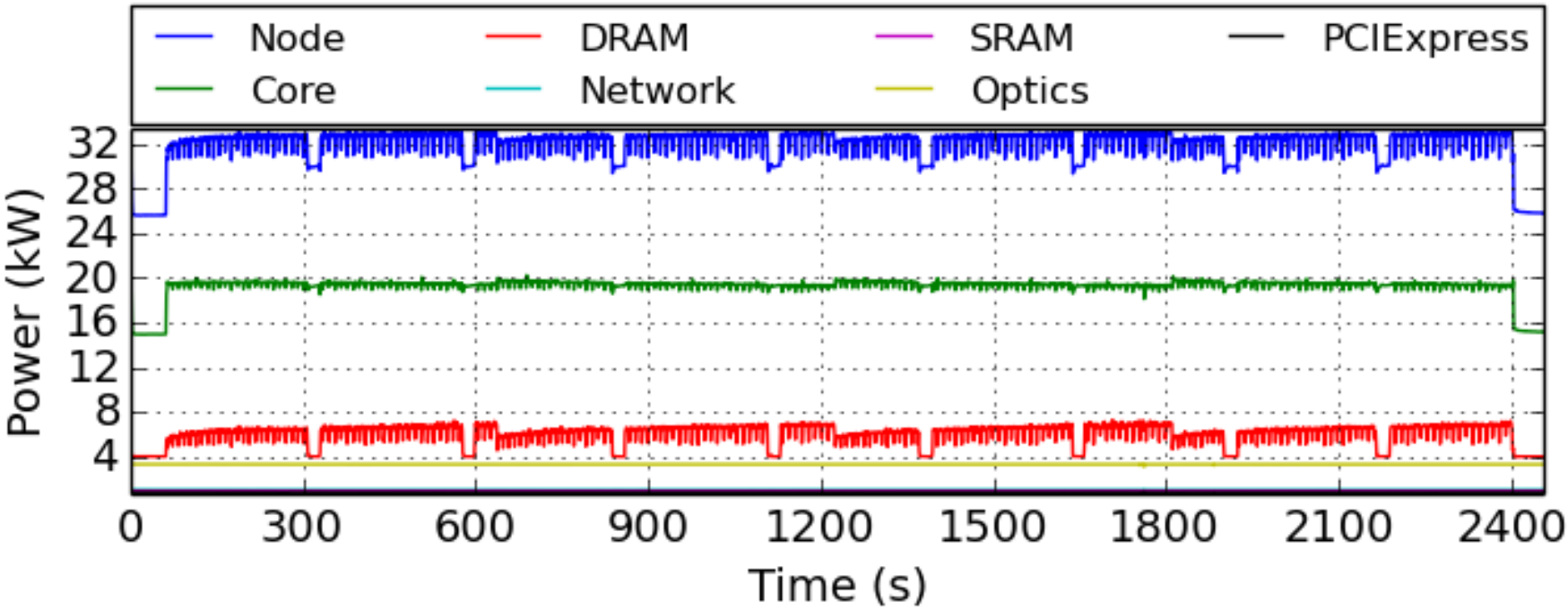
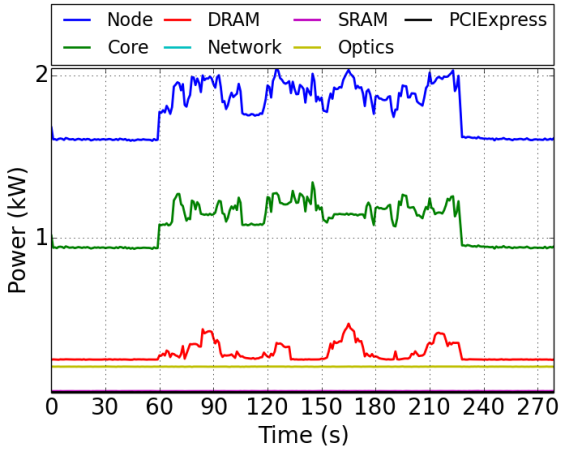


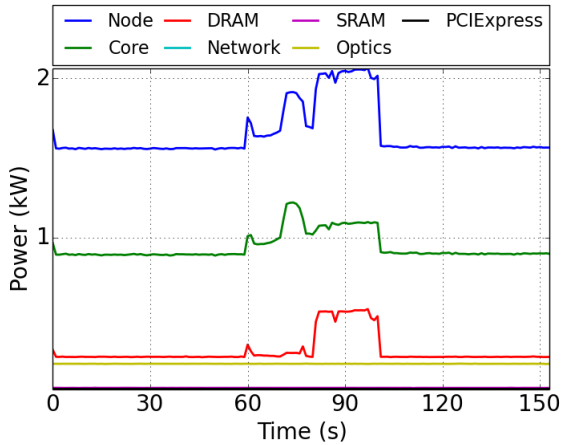
Figure : Asynchronous multilevel checkpointing

1.3 billion atoms Lennard-Jones simulation.
512 nodes running 64 MPI ranks per node (32,678 proc.).

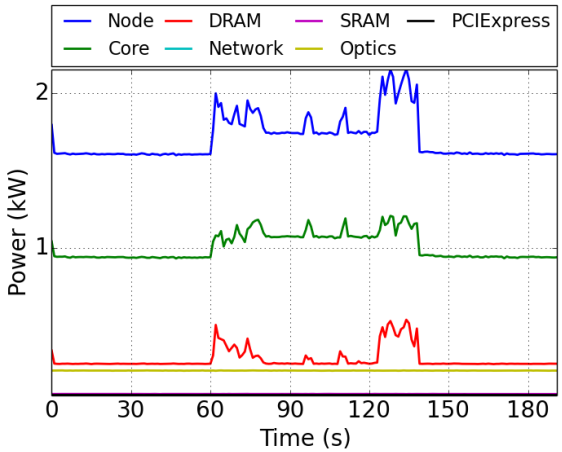
CORAL benchmark



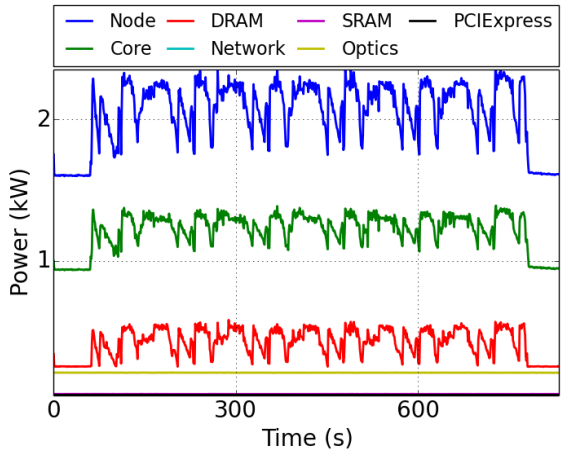
(a) LULESH



(b) MiniFE

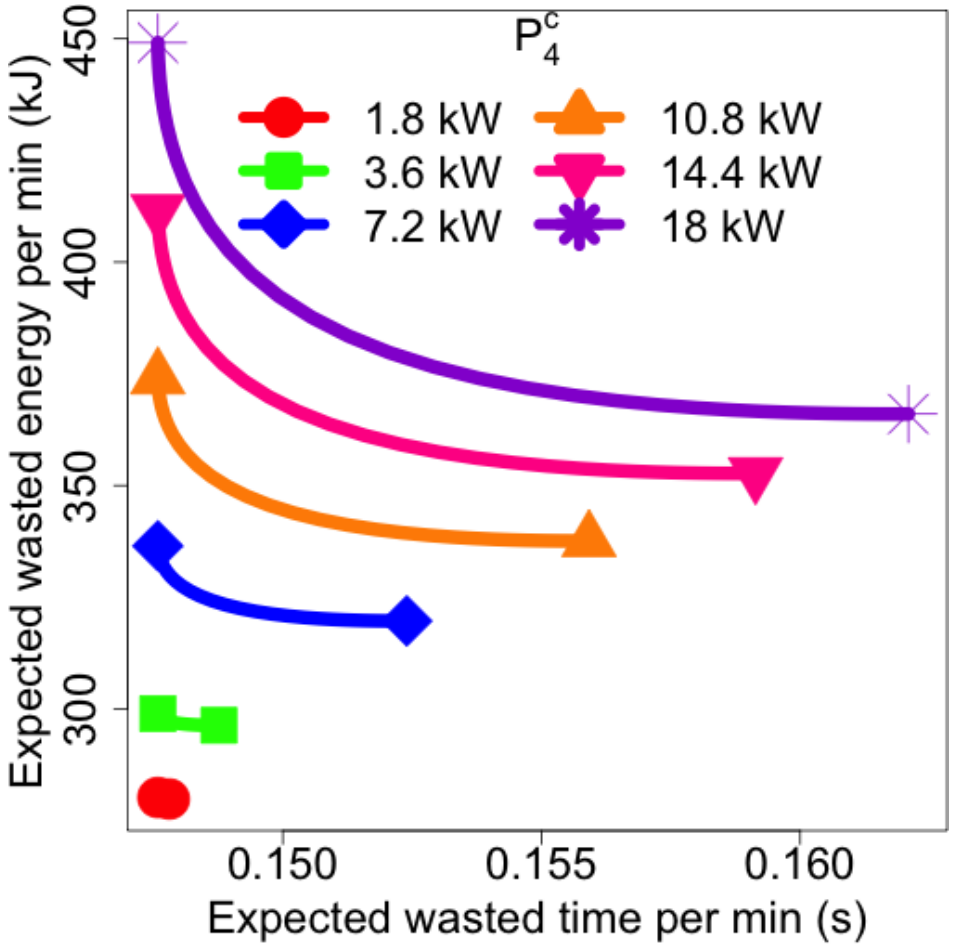


(c) AMG



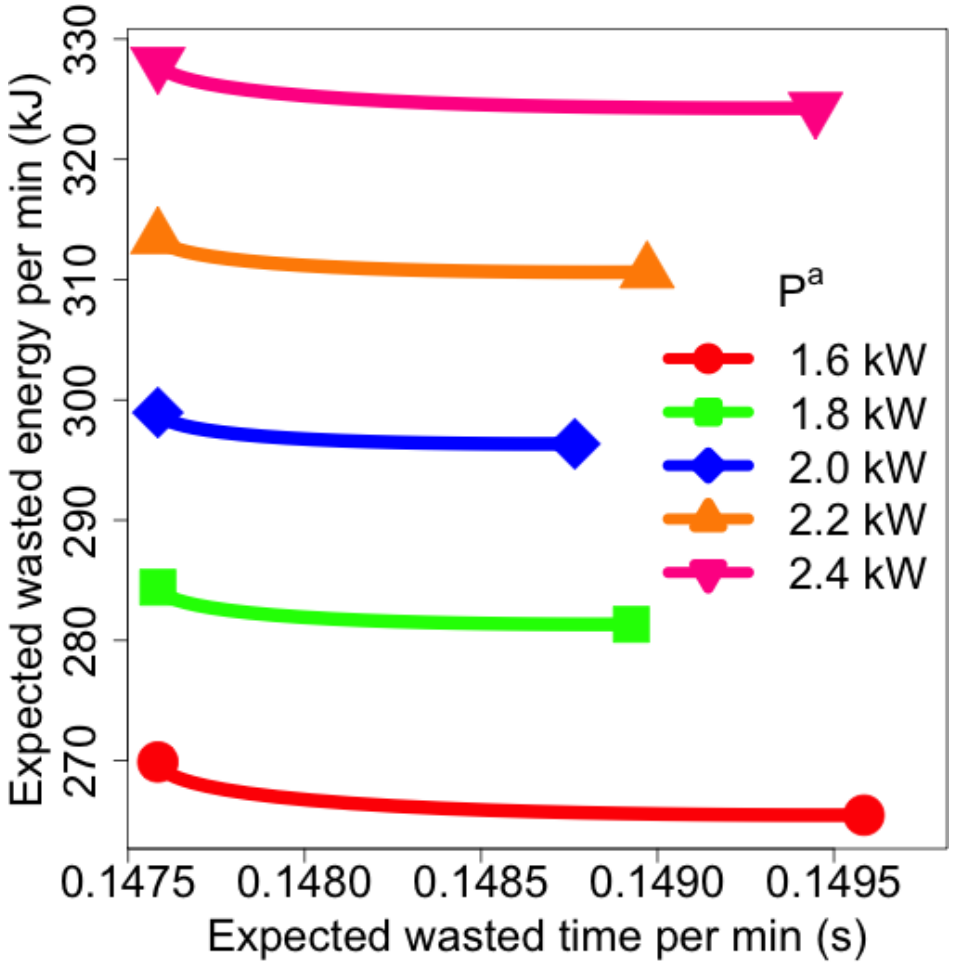
(d) Qbox

Pareto fronts: Level 4 power consumption



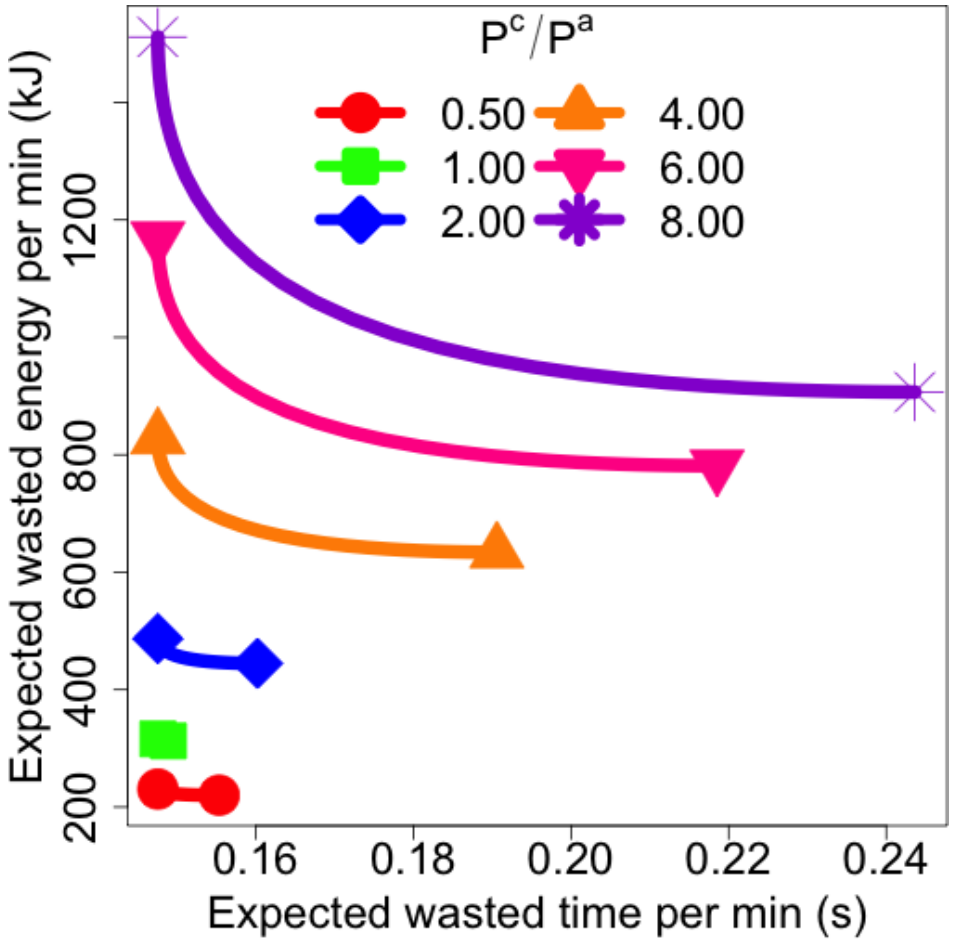
(e) Level 4 power consumption $\mathcal{P}_4^c$

Pareto fronts :Computation power



(f) Computation power $\mathcal{P}^a$

Pareto fronts : Power ratio $\frac{\mathcal{P}^c}{\mathcal{P}^a}$



(g) Power ratio $\frac{\mathcal{P}^c}{\mathcal{P}^a}$

1. Context and motivations
2. Problem formulation and notations
   - Multilevel checkpointing
   - Energy model
   - Multiobjective optimization
3. Simulation and experimentations
   - Experimentations
   - Tradeoff analysis
4. Conclusion and future work

Summary

- Analytical models of performance and energy for multilevel checkpoint schemes.
- The pareto-front is obtained using convex combination.
- Power measurement experiments with production-level scientific applications running on over 32,000 MPI processes:
- The relative energy overhead of using FTI is minor and thus the tradeoffs are relatively small.
- Richer tradeoff exist when the power consumption of checkpointing is greater than that of the computation.

What is Next?

- Analyzing power profile of different fault tolerance protocols such as full/partial replication and message logging.
- The viability of replication with respect to the power cap of future exascale platforms

Questions ?

Thank You !!