Energy Profile of Fault Tolerance Protocols for HPC Systems

Esteban Meneses, Osman Sarood and Sanjay Kalé

Parallel Programming Laboratory
University of Illinois at Urbana-Champaign

Charm++ Workshop, 2013

Exascale

Power & Energy

- Power management (20MW budget)
- Administrative considerations (1MW $\rightarrow$ \$1M/year)
- System codesign (architectural features)

Fault Tolerance

- Size of the machine (200,000 sockets $\rightarrow$ MTBF)
- Types of failures (memory, accelerator, network)
- Different strategies

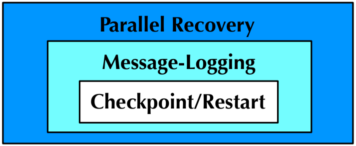
Energy Efficiency of Fault Tolerance Protocols

Agenda

1. Fault Tolerance Protocols
2. Experimental Setup
3. Experimental Results
4. Analytical Model
5. Discussion
6. Conclusions and Future Work

Fault Tolerance Protocols

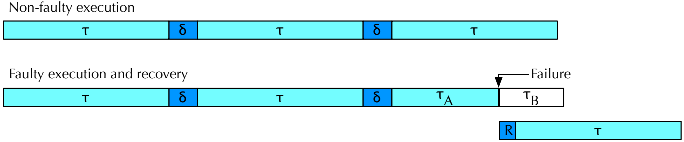
- **Checkpoint/Restart**
  - State is saved periodically
  - Coordinated global checkpoint
  - Checkpoint stored locally
  - Failure $\rightarrow$ global rollback
- **Message-Logging**
  - Messages are stored at sender
  - Non-determinism logged
  - Determinants in causal path
  - Failure $\rightarrow$ local rollback
- **Parallel Recovery**
  - Tasks are migratable
  - Failure $\rightarrow$ recovery in parallel



Caveat

- Many variants of checkpoint/restart
- Several message-logging protocols
- Hybrid schemes

Optimum Checkpoint Period



Daly's modified model:

$$\tau = \sqrt{2\delta(M + R)} - \delta$$

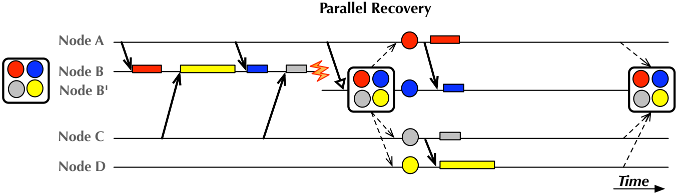
Questions

- Optimum τ for message-logging and parallel recovery?
- Optimum τ to minimize energy?
- Execution time vs energy consumption?

Charm++ Runtime System

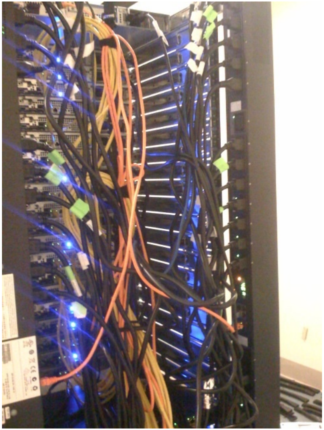
- Migratable Objects Model
- Asynchronous Method Invocation
- Adaptive MPI → each rank becomes an object
- Application-level checkpoint

- One process per *logical* node
- Failure injection: `kill -9 pid`
- Failure detection → automatic restart on replacement node
- Fault tolerance protocols at object-level



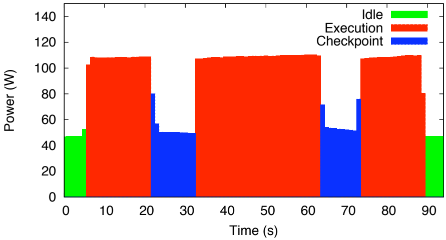
Energy Cluster

- **General Features**
  - 40 single-socket nodes
  - Each node has a four-core Intel Xeon and 4GB of main memory
  - Gigabit ethernet switch
- **Power Measuring**
  - Liebert power distribution unit (PDU)
  - Power measurement per-node
  - 1-second interval frequency



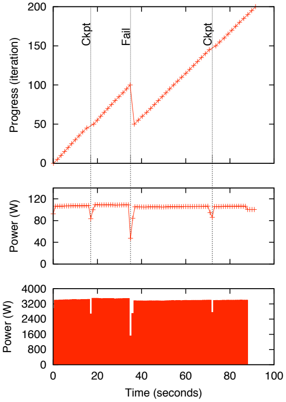
Checkpoint/Restart

- Test program
  - 7-point stencil
  - Nearest neighbor in 3D
  - Barrier after each step
  - Virtualization ratio = 32
  - 200 steps (checkpoints at 50 and 150)
- Local disk checkpoint

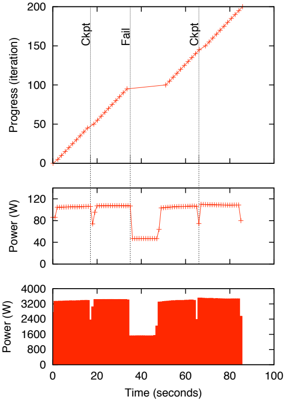


Total Energy Consumed

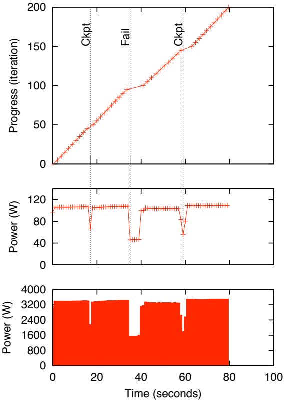
Checkpoint/Restart



Message-Logging

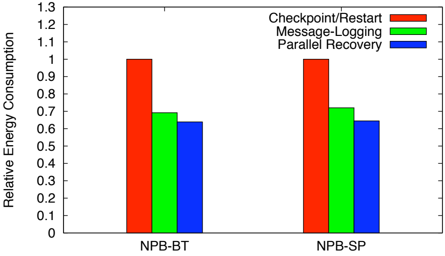


Parallel Recovery



Energy Consumption in Recovery

- Test programs
  - NAS Parallel Benchmarks
  - Block Tridiagonal (BT) and Scalar Pentadiagonal (SP)
  - Virtualization ratio = 4



Summary

	Jacobi3D	NPB-BT	NPB-SP
Language	Charm++	MPI	MPI
Problem size	1024^3	class C	class C
Number of cores	128	100	100
Virtualization ratio	32	4	4
Recovery parallelism	8	4	4
Message-logging overhead	1.0%	3.6%	4.1%
Max power (C)	106	102	95
Max power (M)	106	102	96

Message-logging does NOT increase power draw

Execution Time and Energy Model

Parameter	Description	Value
W	Time to solution with V	24 h
M	Mean-time-to-interrupt of the system	-
δ	Checkpoint time	180 s
τ	Optimum checkpoint period	-
R	Restart time	30 s
T	Total execution time	-
E	Total energy consumption	-
μ	Message-logging slowdown	1.05
P	Available parallelism during recovery	8
ϕ	Message-logging recovery speedup	1.2
σ	Parallel recovery speedup	P
λ	Parallel recovery slowdown	$\frac{P+1}{P}$
H	Max power of each socket	100 W
L	Base power of each socket	40 W

Execution Time Equation

$$T = T_{Solve} + T_{Checkpoint} + T_{Recover} + T_{Restart}$$

Execution Time (Checkpoint/Restart)

$$T = W + \left(\frac{W}{\tau} - 1\right)\delta + \frac{T}{M}\left(\frac{\tau+\delta}{2}\right) + \frac{T}{M}R$$

Execution Time (Parallel Recovery)

$$T = W\mu + \left(\frac{W\mu}{\tau} - 1\right)\delta + \frac{T}{M}\left(\frac{\tau}{\tau+\delta}\left(\frac{\tau}{2\sigma} + \frac{\tau}{2}(\lambda - 1)\right) + \frac{\delta}{\tau+\delta}\left(\frac{\tau}{\sigma} + \frac{\delta}{2}\right)\right) + \frac{T}{M}R$$

Energy Equation

$$E = E_{Solve} + E_{Checkpoint} + E_{Recover} + E_{Restart}$$

Energy (Checkpoint/Restart)

$$E = WSH + \left(\frac{W}{\tau} - 1\right)\delta SL + \frac{T}{M}\left(\frac{\tau}{\tau+\delta} \cdot \frac{\tau}{2} SH + \frac{\delta}{\tau+\delta}\left(\tau SH + \frac{\delta}{2} SL\right)\right) + \frac{T}{M} RSL$$

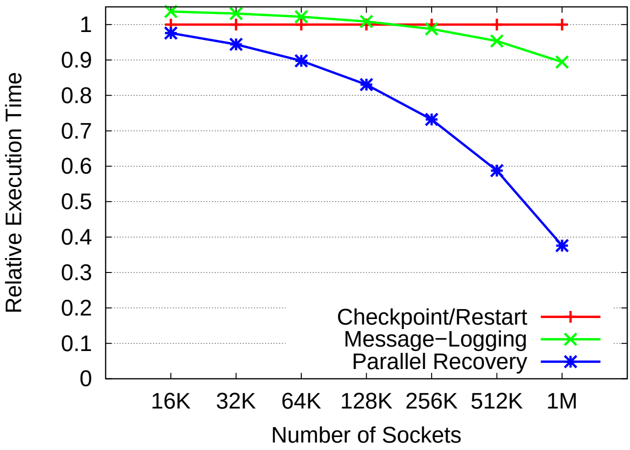
Energy (Parallel Recovery)

$$E = W\mu SH + \left(\frac{W\mu}{\tau} - 1\right)\delta SL +$$
$$\frac{T}{M}\left(\frac{\tau}{\tau+\delta}\left(\frac{\tau}{2\sigma}(PH + (S-P)L) + \frac{\tau}{2}(\lambda - 1)SH\right) + \frac{\delta}{\tau+\delta}\left(\frac{\tau}{\sigma}(PH + (S-P)L) + \frac{\delta}{2} SL\right)\right) + \frac{T}{M} RSL$$

Time-optimum τ

Energy-optimum τ

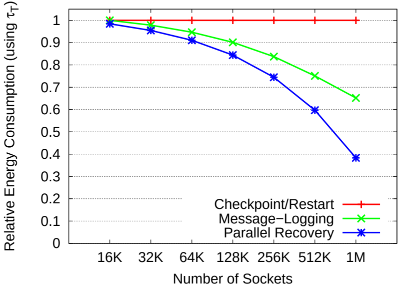
Relative Execution Time



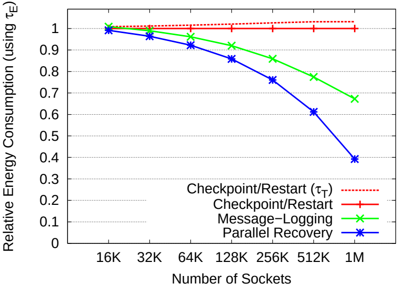
Parallel recovery executes twice as fast

Relative Energy Consumption

Time-optimum τ



Energy-optimum τ

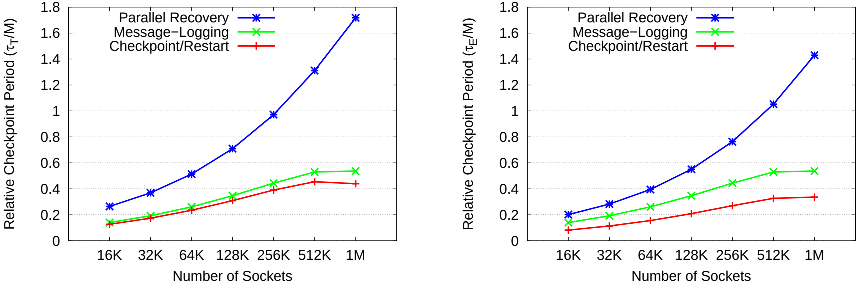


Message-logging consumes 30% less energy

Checkpoint Period

Time-optimum τ

Energy-optimum τ



Parallel recovery checkpoints less often than MTBF

Discussion

- Trend in ratio of base to maximum power

Processor	Release Date	Max Power	Base Power	Base/Max Ratio
Intel Xeon (E5520)	Q1,09	125	60	0.48
Intel Nehalem (i7 860)	Q3,09	151	52	0.34
Intel Sandy Bridge (i7 2600)	Q1,11	101	21	0.21

- Migratability and over-decomposition in scientific applications

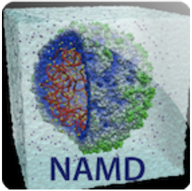
Conclusions

- *"Minimize execution time $\implies$ minimize energy"* (not true)
  - Increase checkpoint frequency
  - Recovery is more energy-efficient with message logging
- Energy overhead of message-logging
  - It does not increase power draw
  - It increases energy consumption on the forward path
- Parallel recovery leverages message-logging
  - It provides the minimum execution time (users happy)
  - It offers the minimum energy consumed (administrators happy)
  - The model predicts more than **50%** reduction in execution time and more than **50%** reduction in energy consumed at extreme scale

Future Work

- Particle-simulation applications:

Molecular Dynamics



Quantum Chemistry

Cosmology

- Enhancements to analytical model:
  - Different failure distributions: Weibull, log-normal
  - No upper bound for checkpoint period
- Energy-aware fault tolerance protocols

Acknowledgements

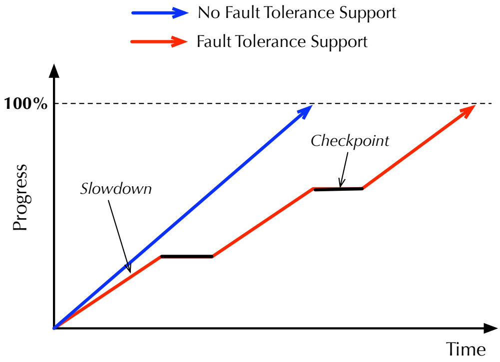
- **HPC Colony II Project**. This work is partially supported by the US Department of Energy under grant DOE DE-SC0001845 and by a machine allocation on XSEDE under award ASC050039N.
- **Prof. Tarek F. Abdelzaher**. The experimental results of this work come from the *Energy Cluster* in the University of Illinois at Urbana-Champaign.

Thank you!

Q&A

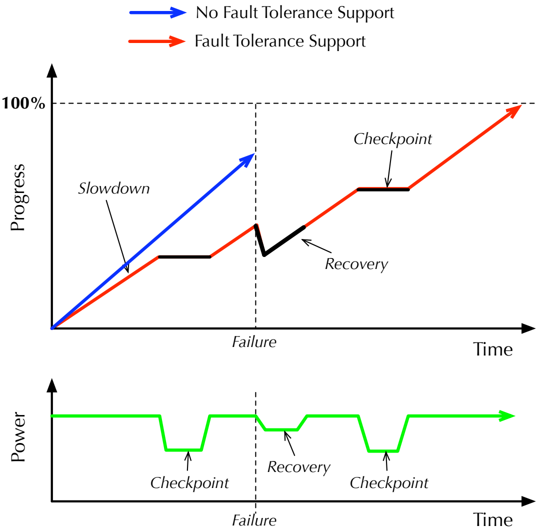
11th Annual Workshop on Charm++ and its Applications

Progress Diagram

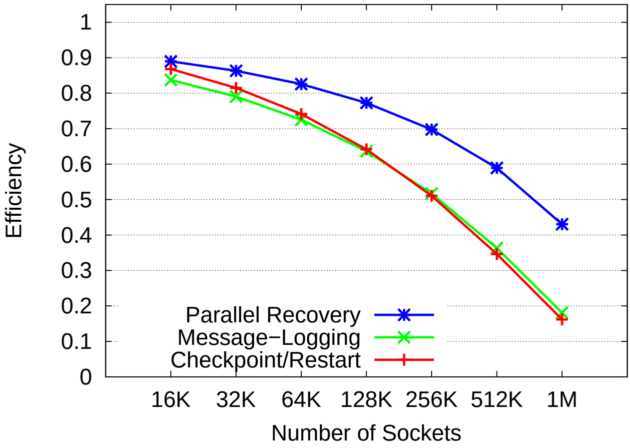


Performance Overhead

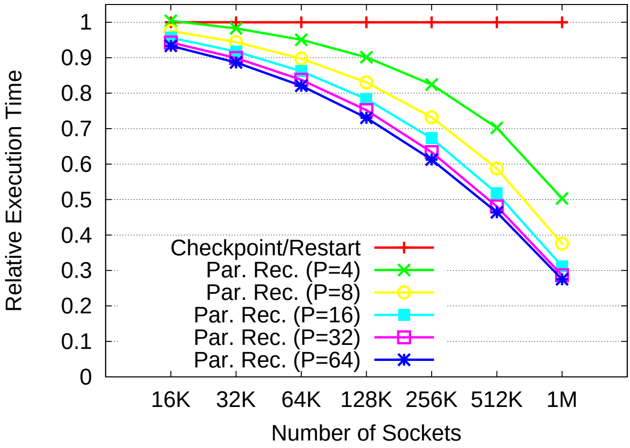
Progress Diagram for Energy Efficient Fault Tolerance



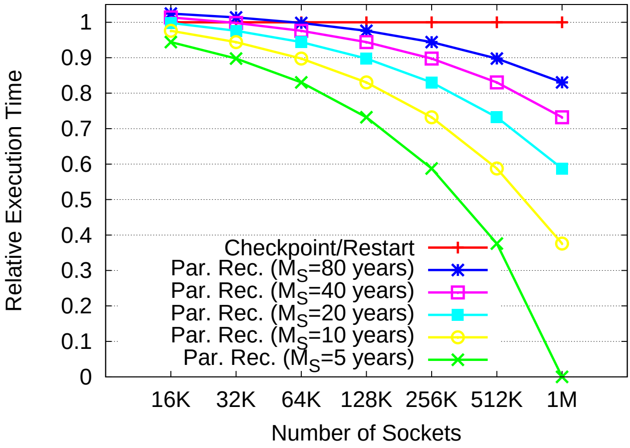
System Utilization



Effect of Higher Parallelism During Recovery



Effect of Failure Rate per Socket



Simulation Results