Increasing Fault Resiliency in a Message-Passing Environment

Rolf Riesen, Kurt Ferreira
Ron Oldfield, Jon Stearley
James Laros, Kevin Pedretti
Ron Brightwell
rolf@sandia.gov
Sandia National Laboratories

Todd Kordenbrock
Hewlett-Packard Company

October 14, 2009

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.

Motivation
Design
Evaluation
Analysis
Implications
Summary

Motivation

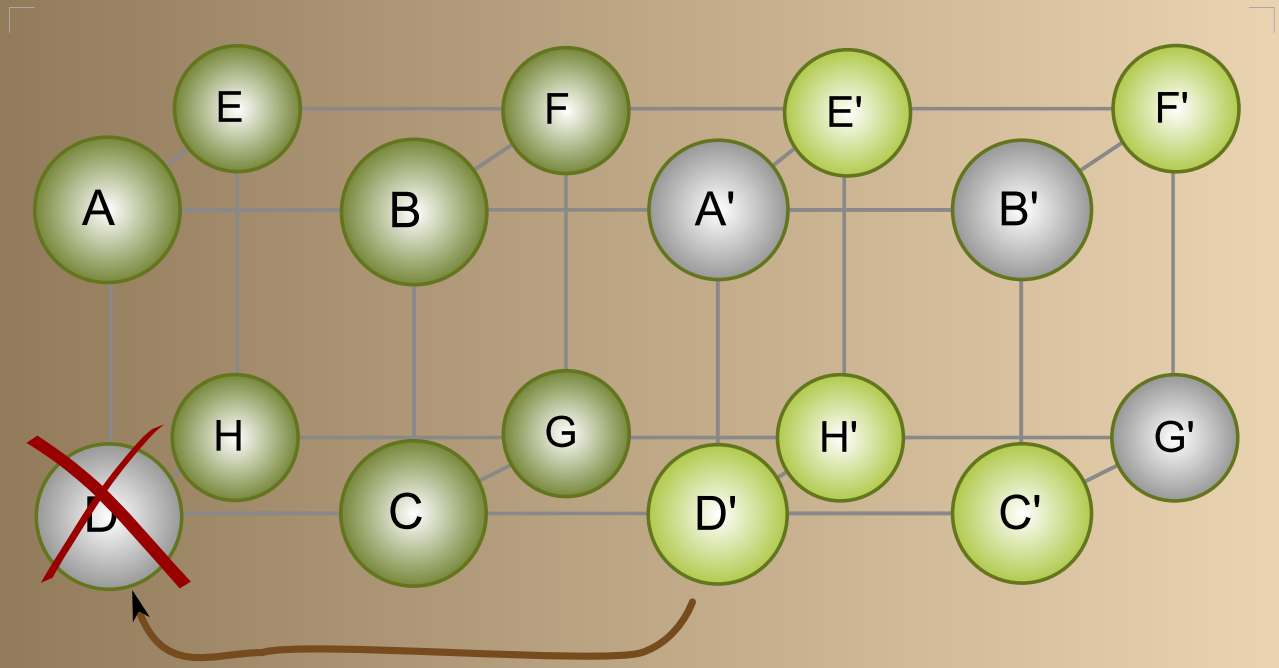
Motivation
Design
Evaluation
Analysis
Implications
Summary

- Checkpoint/Restart is common way to deal with faults
- What can we do to increase checkpoint interval?
- Explore redundant computation for MPI applications
  - Can it be done at user level?
  - What are requirements for RAS system and runtime?
  - What is the overhead
  - Cost versus benefit?
- Write *r*MPI library at MPI profiling layer to learn what the issues are

Motivation
Design
Redundancy
Basics
Msg. order
Other
Status
Evaluation
Analysis
Implications
Summary

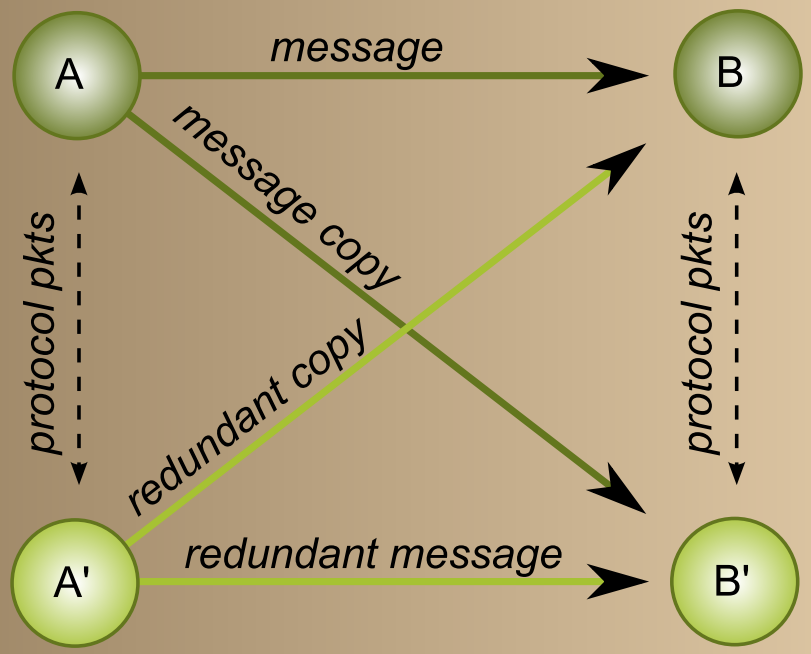
Design

Motivation
Design
Redundancy
Basics
Msg. order
Other
Status
Evaluation
Analysis
Implications
Summary



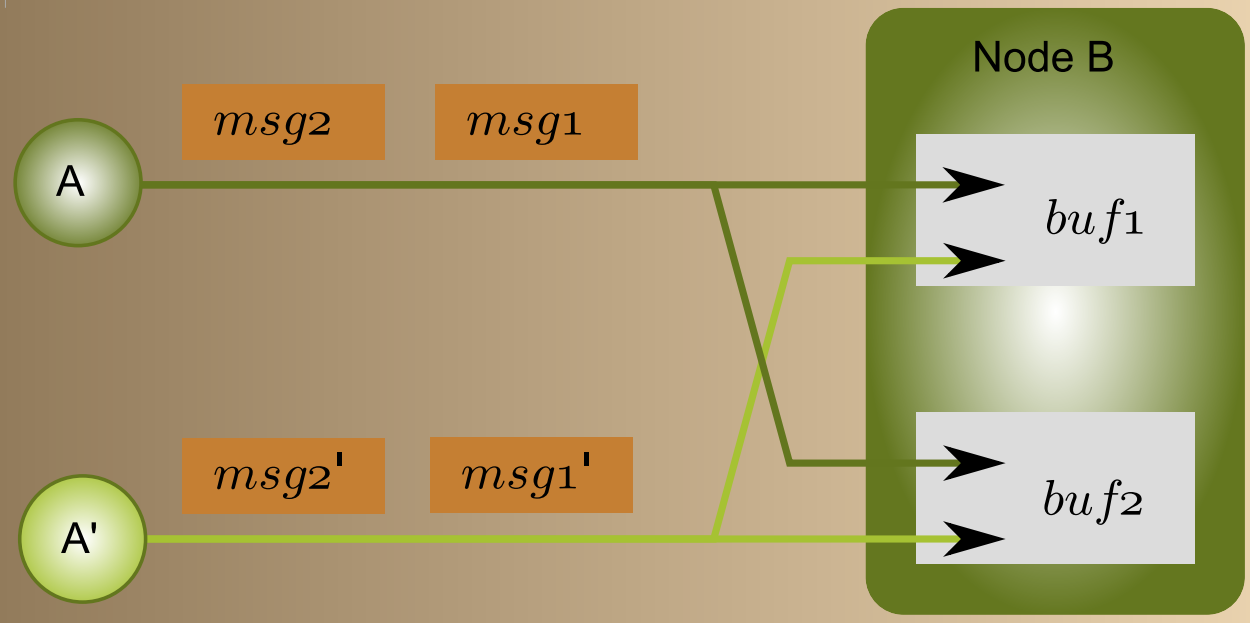
- Active and redundant node do same computation
- One node continues when the other fails
- `MPI_Comm_rank()` returns same value on both nodes
- Not each node needs to have a redundant partner

Motivation
Design
Redundancy
Basics
Msg. order
Other
Status
Evaluation
Analysis
Implications
Summary



- Each message gets sent twice (4, if we count redundant nodes)
- Msg and redundant copy received into same buffer
- Redundant msg have unused tag bit set
- Protocol needed to coordinate receives and other MPI ops

Motivation
Design
Redundancy
Basics
Msg. order
Other
Status
Evaluation
Analysis
Implications
Summary



- Active and redundant node must receive msg in same order
- `MPI_ANY_SOURCE` and `MPI_ANY_TAG` are problematic
- Redundant node maintains queue of posted receives if `MPI_ANY_SOURCE`
- Coordinate with active node to post specific receive

Other Functions

Motivation
Design
Redundancy
Basics
Msg. order
Other
Status
Evaluation
Analysis
Implications
Summary

- Redundant node must return same info for probe, test, and time functions
- Active node does operation and sends result to redundant node
- Collectives use redundant point-to-point
- *r*MPI re-implements almost all of MPI
  - *r*MPI uses MPICH (mostly) as a transport layer

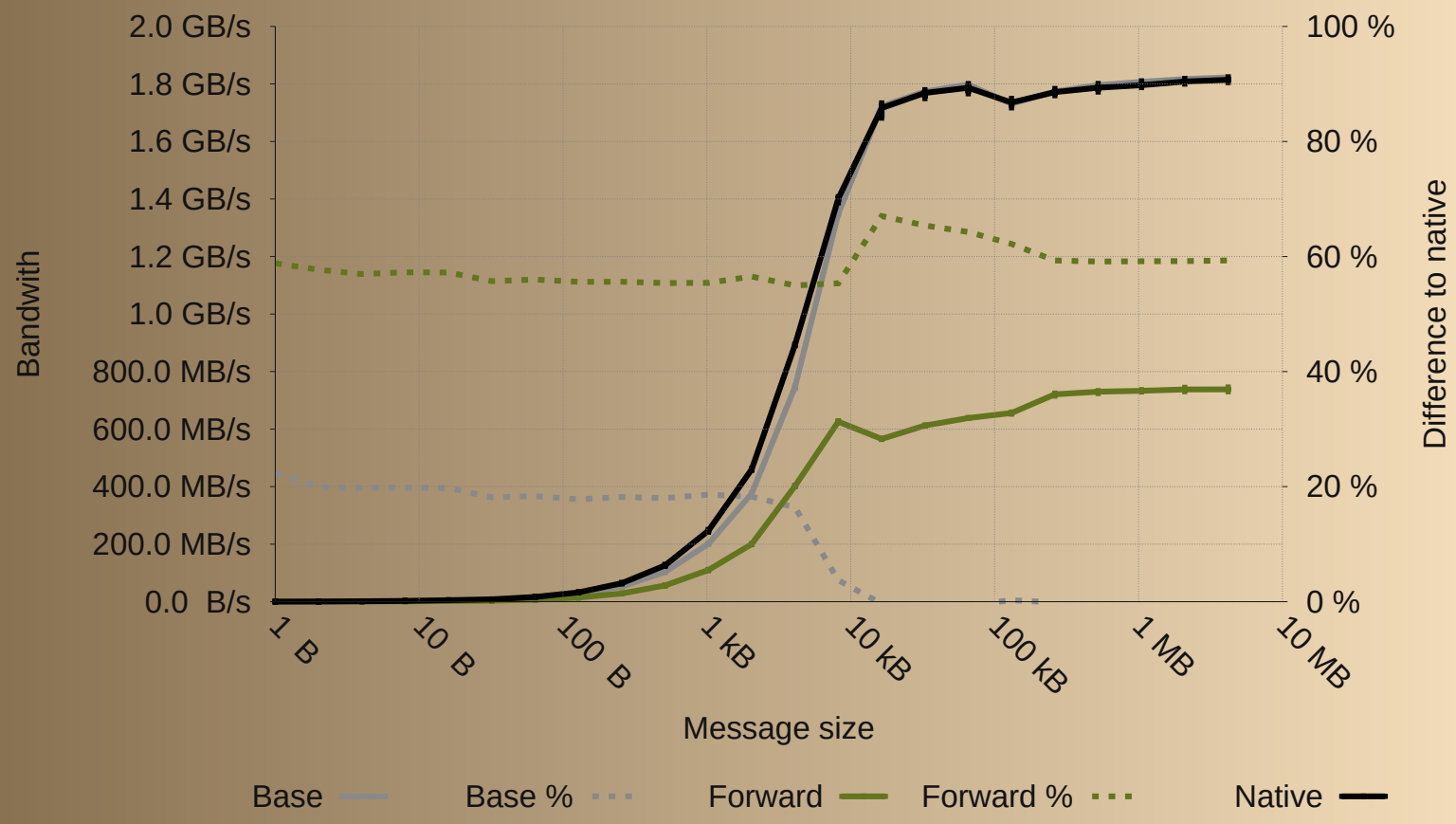
Motivation
Design
Redundancy
Basics
Msg. order
Other
Status
Evaluation
Analysis
Implications
Summary

- Can't do `MPI_ANY_TAG` and `MPI_ANY_SOURCE` simultaneously
- Most mayor functions of MPI-2 implemented
- Are considering transactions to limit data volume
- Plan to open source it
- Few RAS features needed:
  - Notification of node availability
  - No Byzantine behavior (error correction protocol on network)
  - Messages to/from dead nodes must be consumed (no dead- or life-lock)

Motivation
Design
Evaluation
Bandwidth
Latency
Allreduce
CTH
SAGE
LAMMPS
HPCCG
Analysis
Implications
Summary

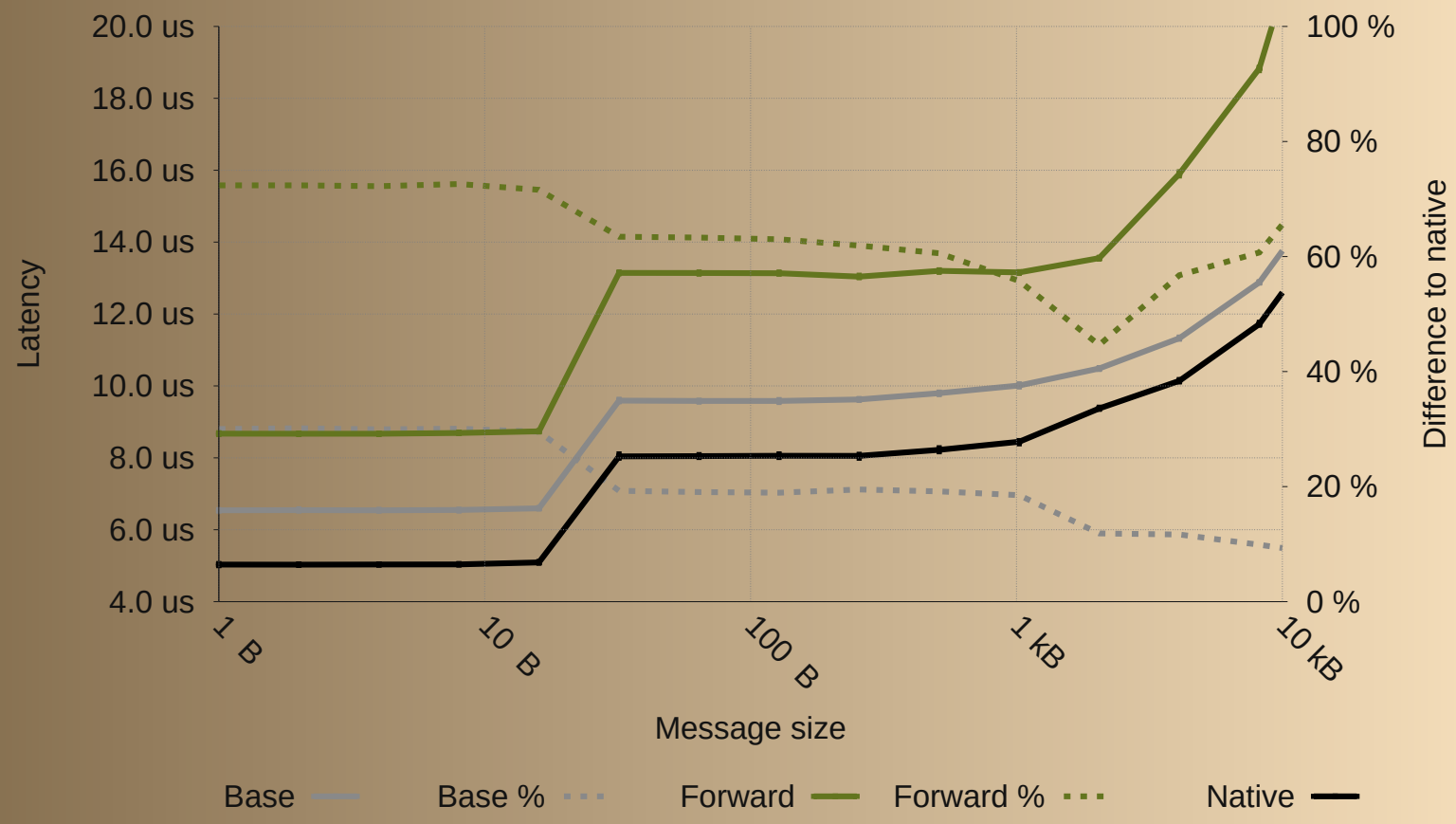
Evaluation

Motivation
Design
Evaluation
Bandwidth
Latency
Allreduce
CTH
SAGE
LAMMPS
HPCCG
Analysis
Implications
Summary



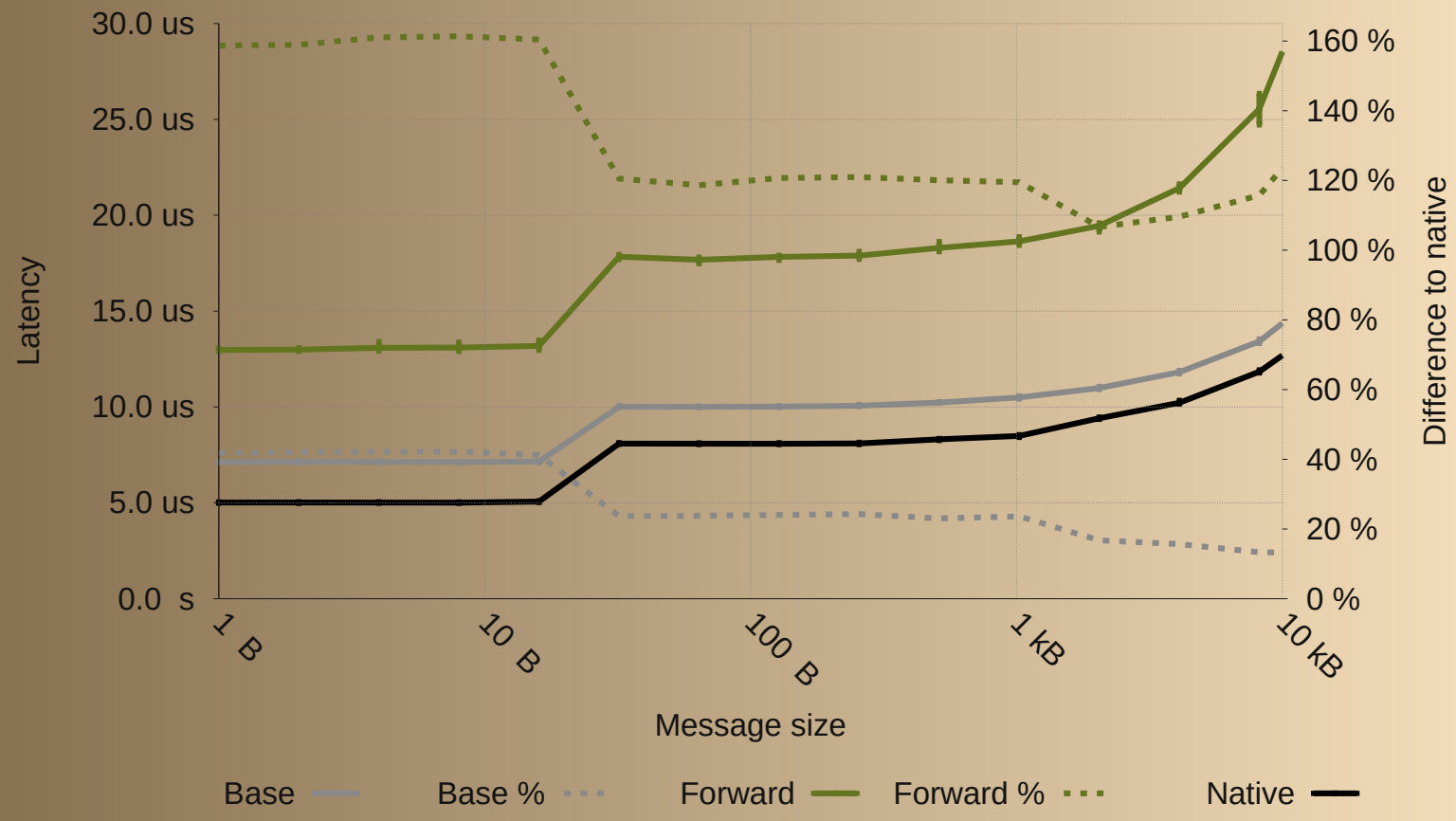
Native	Benchmark w/o *r*MPI
Base	*r*MPI, no redundancy
Forward	ABCD|A'B'C'D'

Motivation
Design
Evaluation
Bandwidth
Latency
Allreduce
CTH
SAGE
LAMMPS
HPCCG
Analysis
Implications
Summary



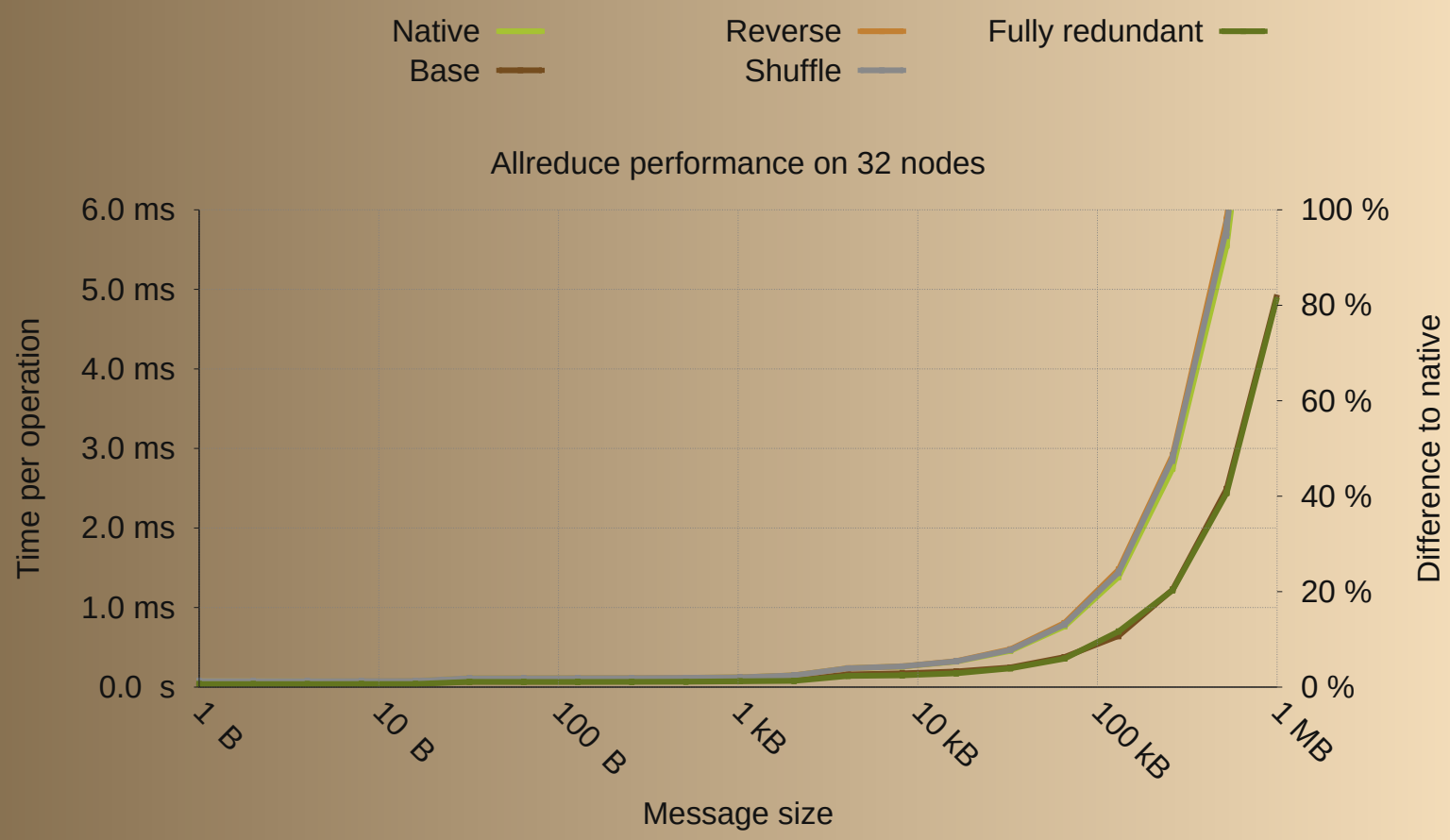
Native	Benchmark w/o *r*MPI
Base	*r*MPI, no redundancy
Forward	ABCD|A'B'C'D'

Motivation
Design
Evaluation
Bandwidth
Latency
Allreduce
CTH
SAGE
LAMMPS
HPCCG
Analysis
Implications
Summary



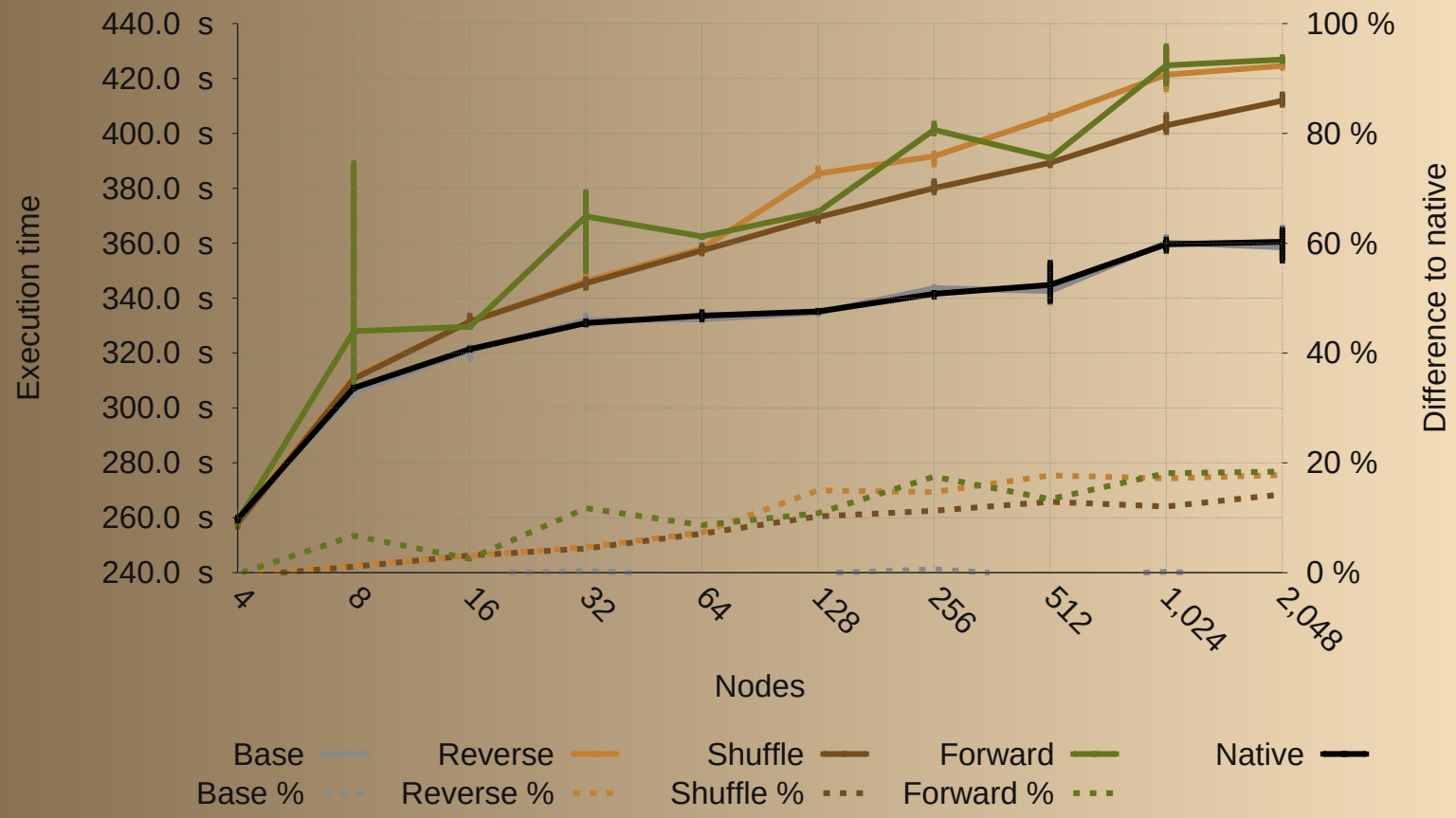
Native	Benchmark w/o *r*MPI
Base	*r*MPI, no redundancy
Forward	ABCD|A'B'C'D'

Motivation
Design
Evaluation
Bandwidth
Latency
Allreduce
CTH
SAGE
LAMMPS
HPCCG
Analysis
Implications
Summary



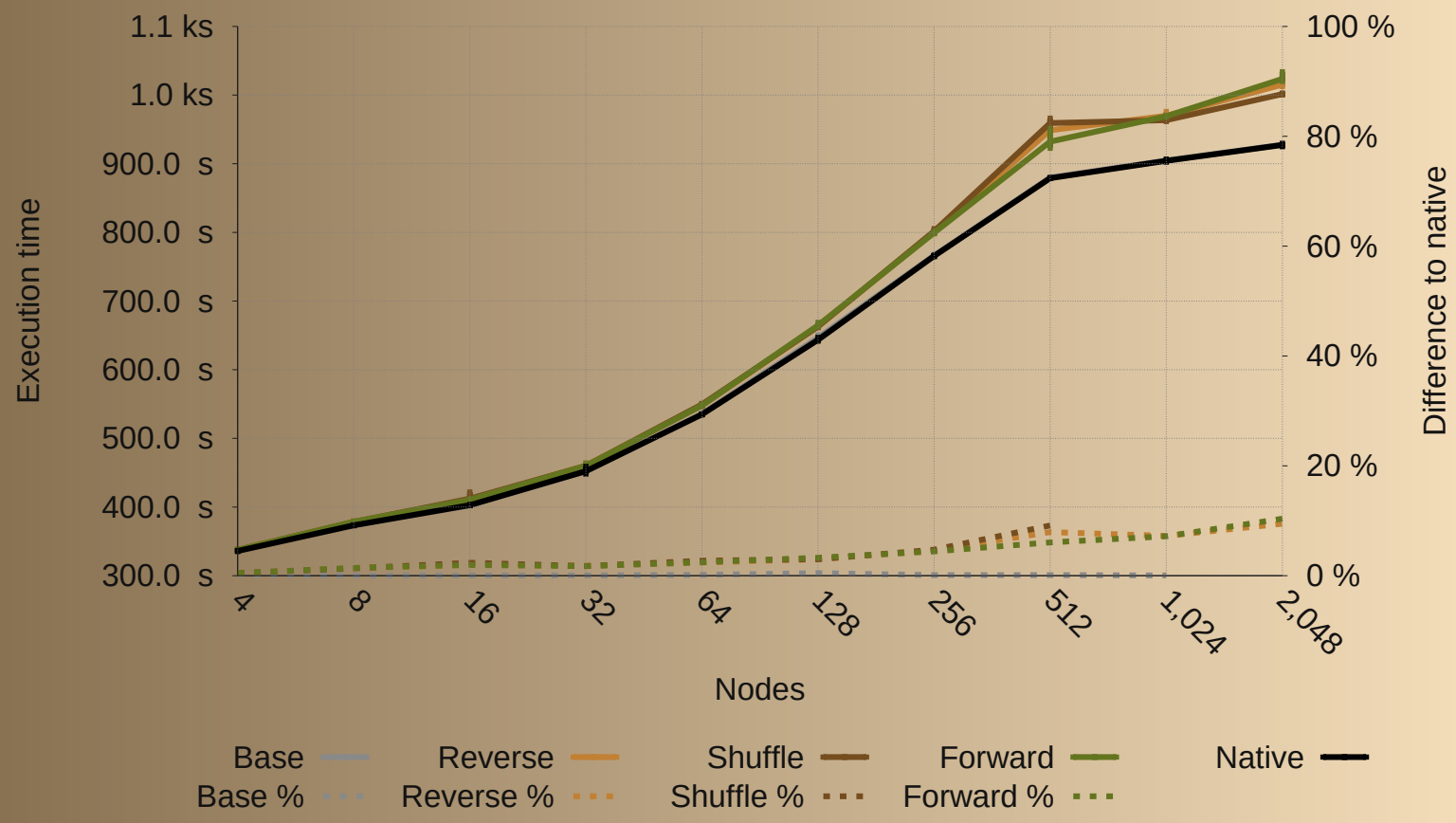
Native	Benchmark w/o *r*MPI		
Base	*r*MPI, no redundancy	Reverse	ABCD|D'C'B'A'
Forward	ABCD|A'B'C'D'	Shuffle	e.g., ABCD|C'B'D'A'

Motivation
Design
Evaluation
Bandwidth
Latency
Allreduce
CTH
SAGE
LAMMPS
HPCCG
Analysis
Implications
Summary



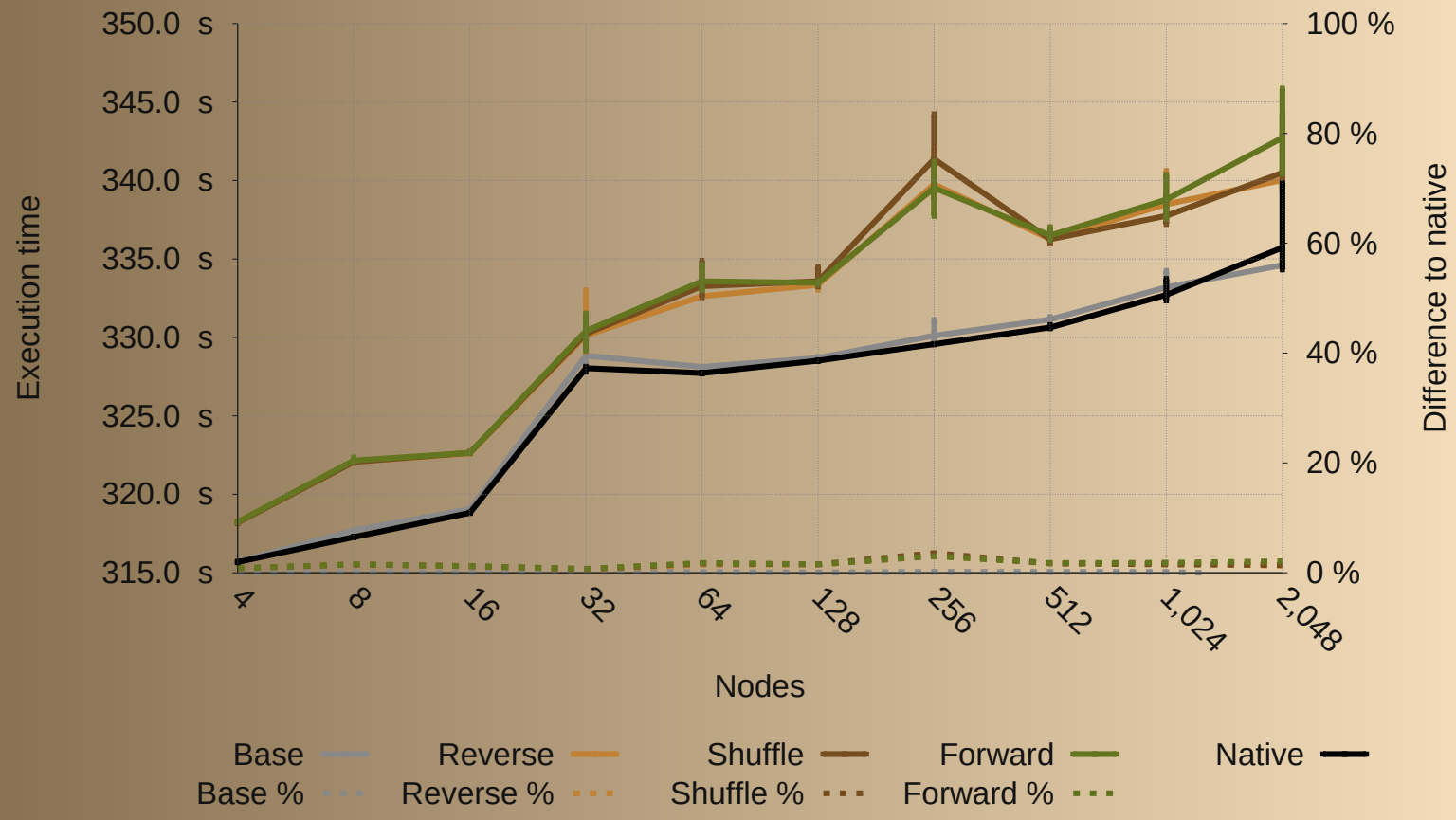
Native	Benchmark w/o *r*MPI		
Base	*r*MPI, no redundancy	Reverse	ABCD|D'C'B'A'
Forward	ABCD|A'B'C'D'	Shuffle	e.g., ABCD|C'B'D'A'

Motivation
Design
Evaluation
Bandwidth
Latency
Allreduce
CTH
SAGE
LAMMPS
HPCCG
Analysis
Implications
Summary



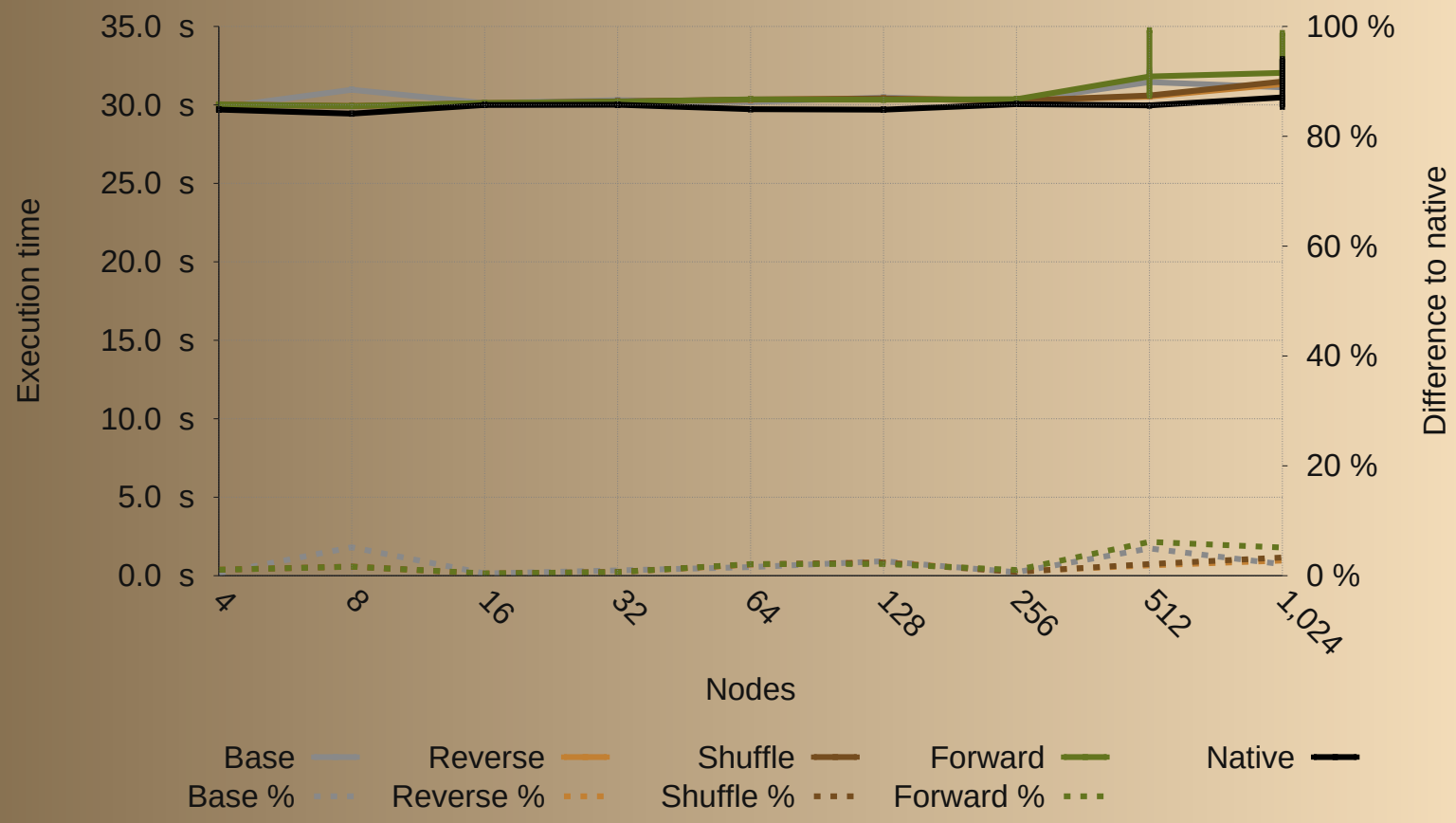
Native	Benchmark w/o *r*MPI		
Base	*r*MPI, no redundancy	Reverse	ABCD|D'C'B'A'
Forward	ABCD|A'B'C'D'	Shuffle	e.g., ABCD|C'B'D'A'

Motivation
Design
Evaluation
Bandwidth
Latency
Allreduce
CTH
SAGE
LAMMPS
HPCCG
Analysis
Implications
Summary



Native	Benchmark w/o *r*MPI		
Base	*r*MPI, no redundancy	Reverse	ABCD|D'C'B'A'
Forward	ABCD|A'B'C'D'	Shuffle	e.g., ABCD|C'B'D'A'

Motivation
Design
Evaluation
Bandwidth
Latency
Allreduce
CTH
SAGE
LAMMPS
HPCCG
Analysis
Implications
Summary

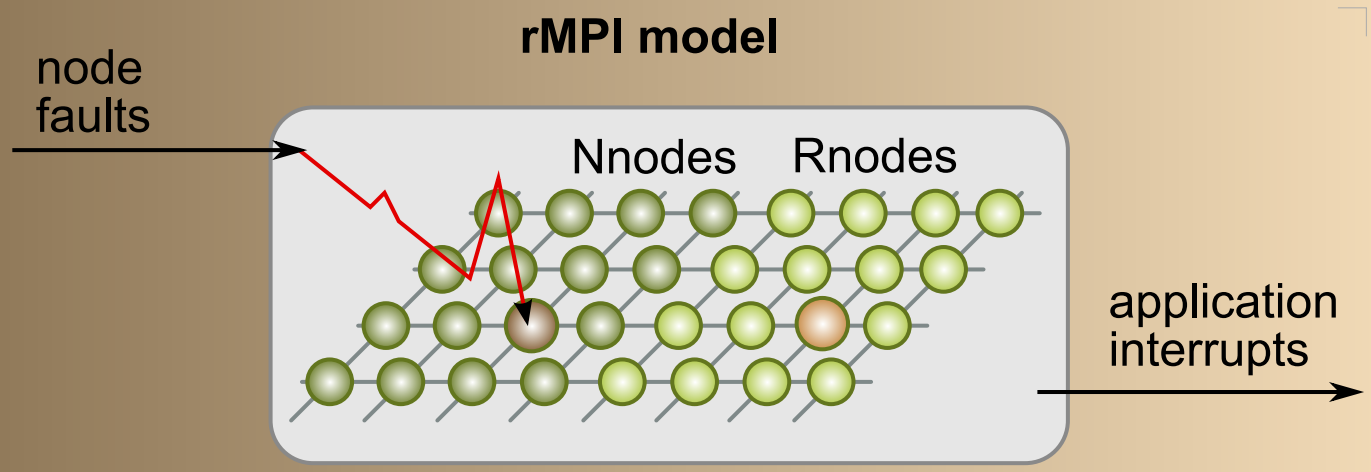


Native	Benchmark w/o *r*MPI		
Base	*r*MPI, no redundancy	Reverse	ABCD|D'C'B'A'
Forward	ABCD|A'B'C'D'	Shuffle	e.g., ABCD|C'B'D'A'

Motivation
Design
Evaluation
Analysis
*r*MPI Model
Interrupts
Lifetime
Simulation
Behavior
Validation
Implications
Summary

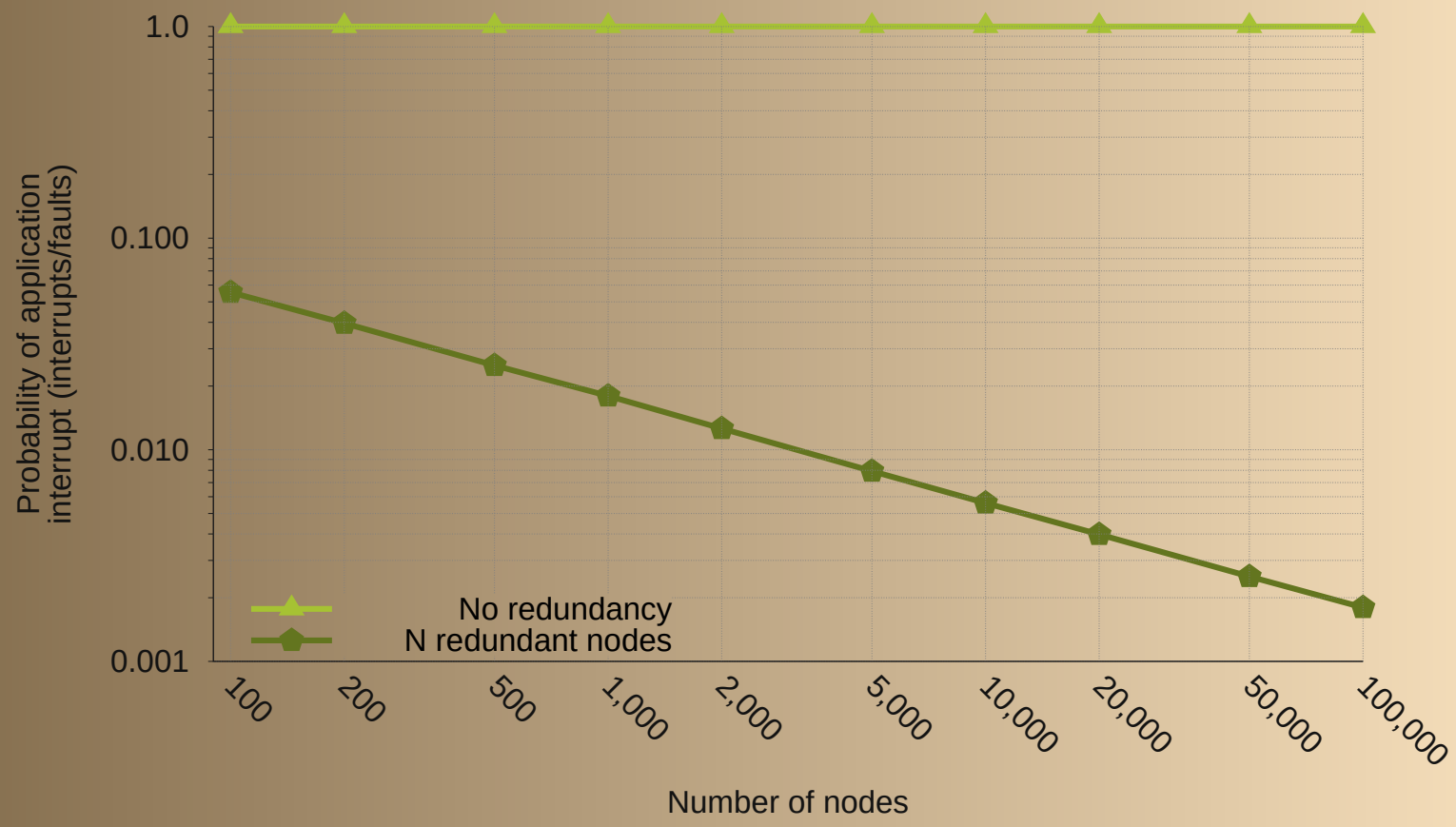
Analysis

Motivation
Design
Evaluation
Analysis
*r*MPI Model
Interrupts
Lifetime
Simulation
Behavior
Validation
Implications
Summary

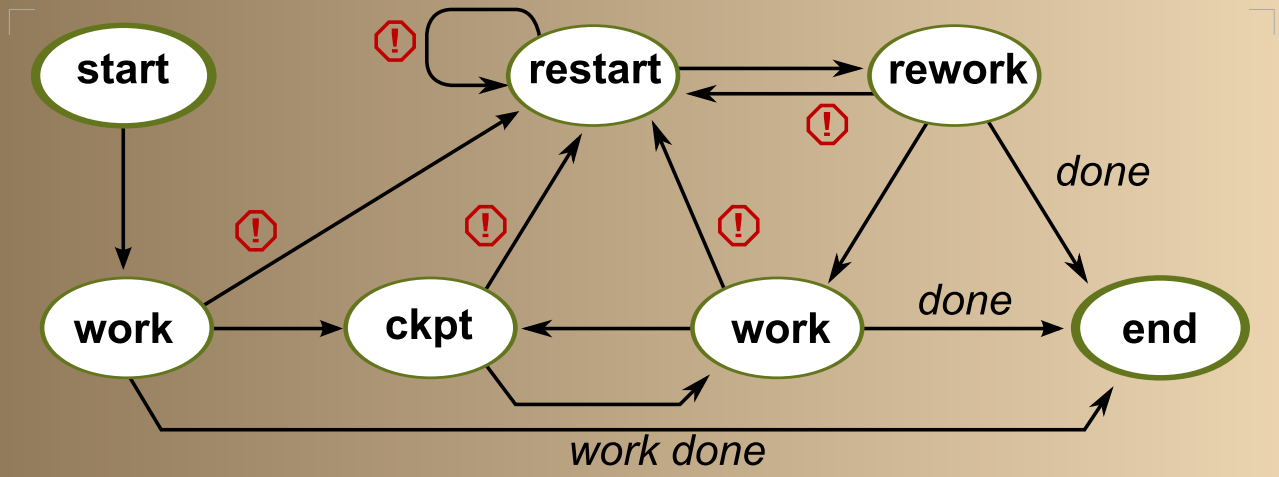


- Acts as a filter
  - Input is a series of faults
  - Output is application interrupts that cause a restart

Motivation
Design
Evaluation
Analysis
*r*MPI Model
Interrupts
Lifetime
Simulation
Behavior
Validation
Implications
Summary

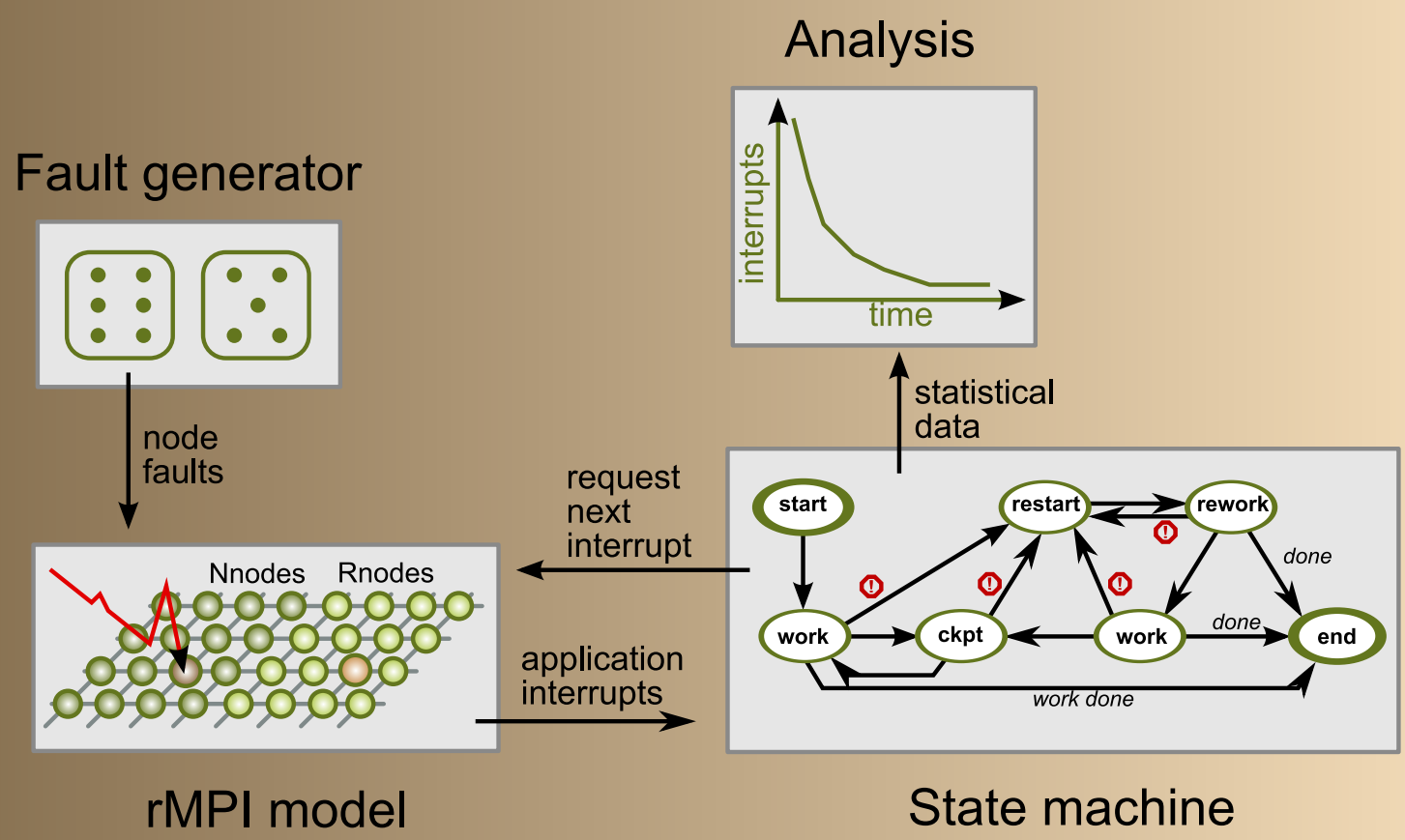


Motivation
Design
Evaluation
Analysis
*r*MPI Model
Interrupts
Lifetime
Simulation
Behavior
Validation
Implications
Summary



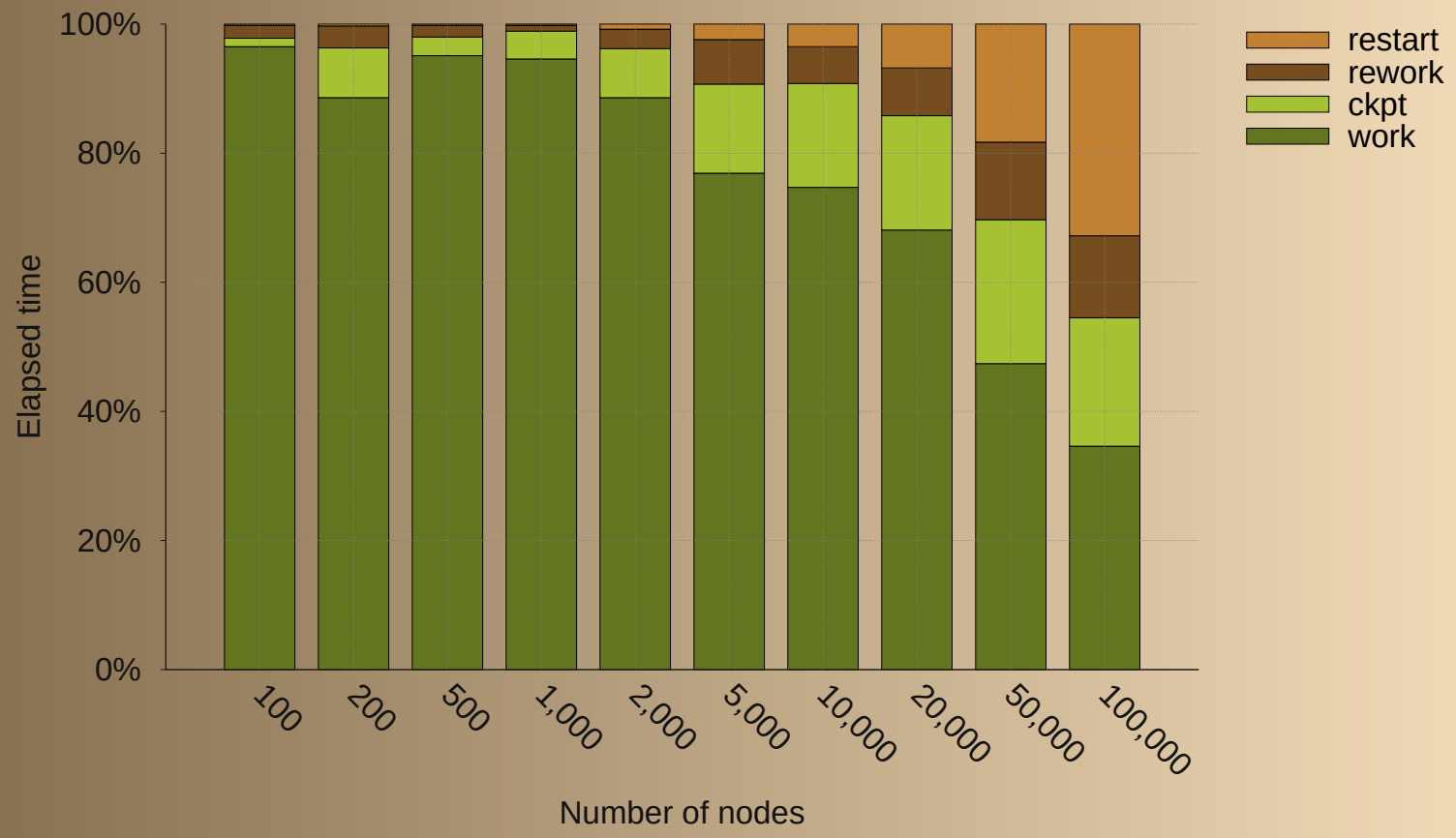
- Application needs to finish a set amount of work
- Do checkpoints and restart when necessary

Motivation
Design
Evaluation
Analysis
*r*MPI Model
Interrupts
Lifetime
Simulation
Behavior
Validation
Implications
Summary



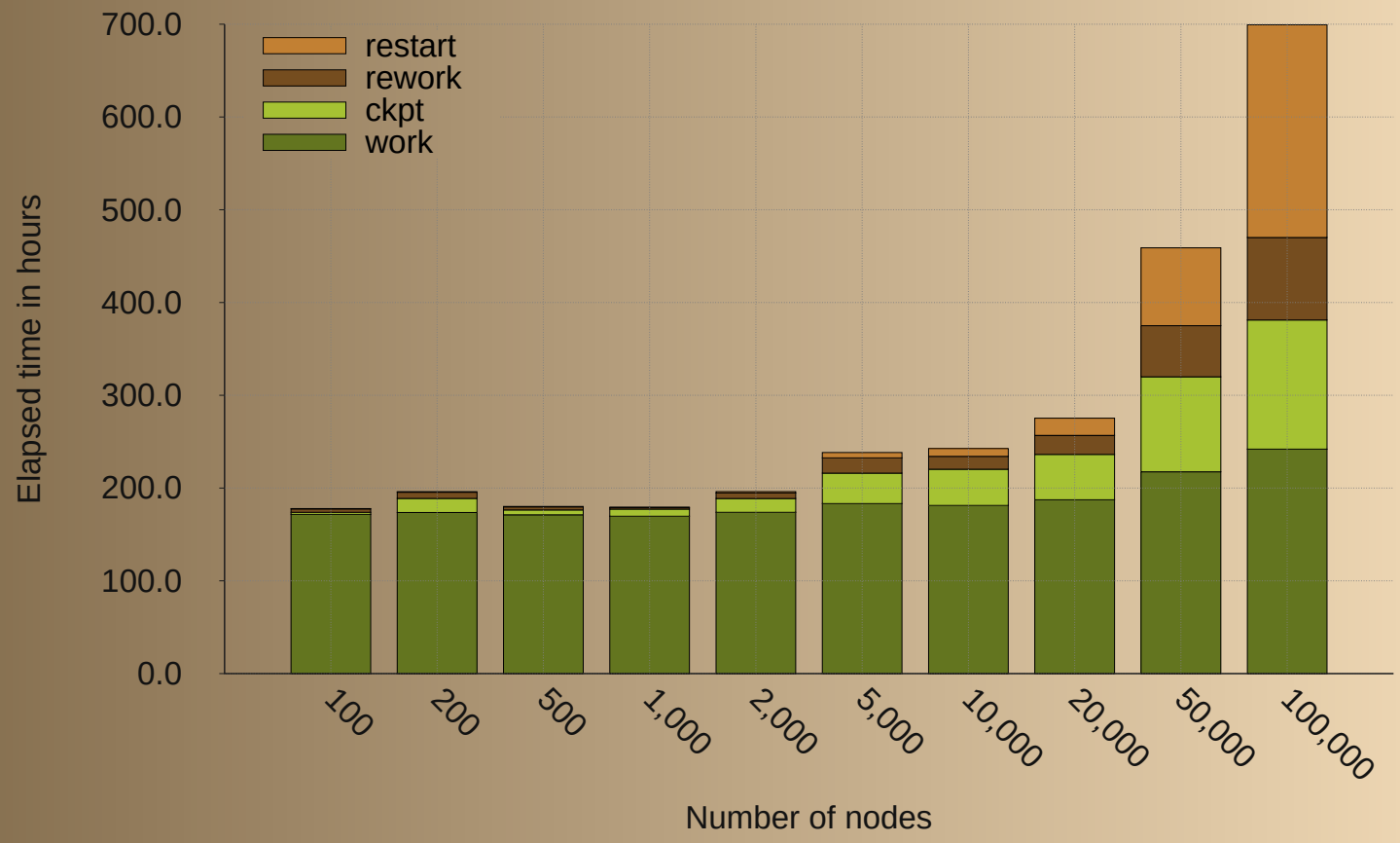
- Combine *r*MPI model with state machine
- Finish a set amount of work
- Do checkpoints and restart when necessary

Motivation
Design
Evaluation
Analysis
*r*MPI Model
Interrupts
Lifetime
Simulation
Behavior
Validation
Implications
Summary



Normalized time spent in work, checkpoint, restart, and rework.
168h work, 5 min checkpoint, 10 min restart, 5-year node MTBF

Motivation
Design
Evaluation
Analysis
*r*MPI Model
Interrupts
Lifetime
Simulation
Behavior
Validation
Implications
Summary



Total time spent in work, checkpoint, restart, and rework.
168h work, 5 min checkpoint, 10 min restart, 5-year node MTBF

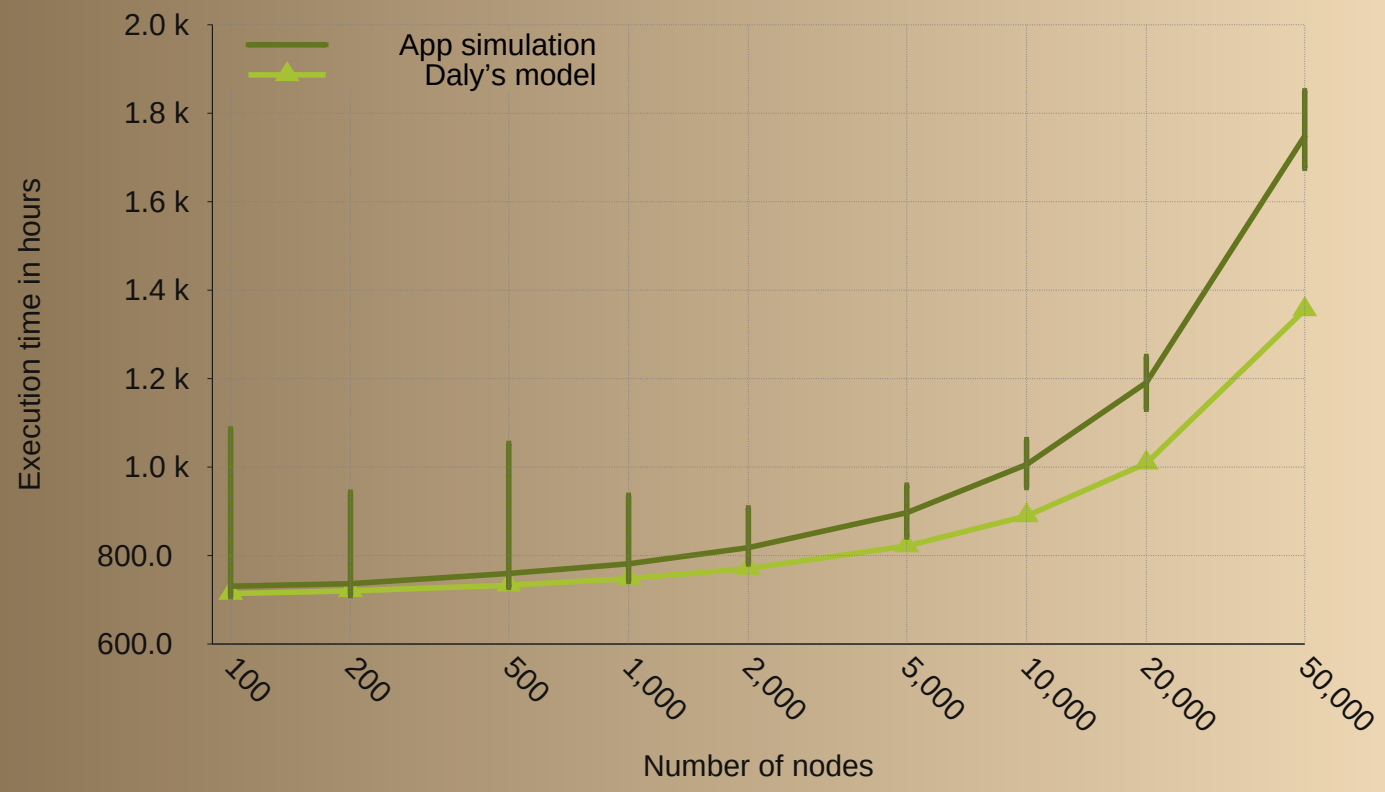
Motivation
Design
Evaluation
Analysis
*r*MPI Model
Interrupts
Lifetime
Simulation
Behavior
Validation
Implications
Summary

Normalized time spent in work, checkpoint, restart, and rework.
168h work, 5 min checkpoint, 10 min restart, 5-year node MTBF

Motivation
Design
Evaluation
Analysis
*r*MPI Model
Interrupts
Lifetime
Simulation
Behavior
Validation
Implications
Summary

Daly's equation from *A higher order estimate of the optimum checkpoint interval for restart dumps*.

$$T_w(\tau) = \Theta e^{\frac{R}{\Theta}}(e^{\frac{\tau+\delta}{\Theta}} - 1)\frac{T_s}{\tau} \quad \text{for} \quad \delta << T_s$$



Motivation
Design
Evaluation
Analysis
Implications
Summary

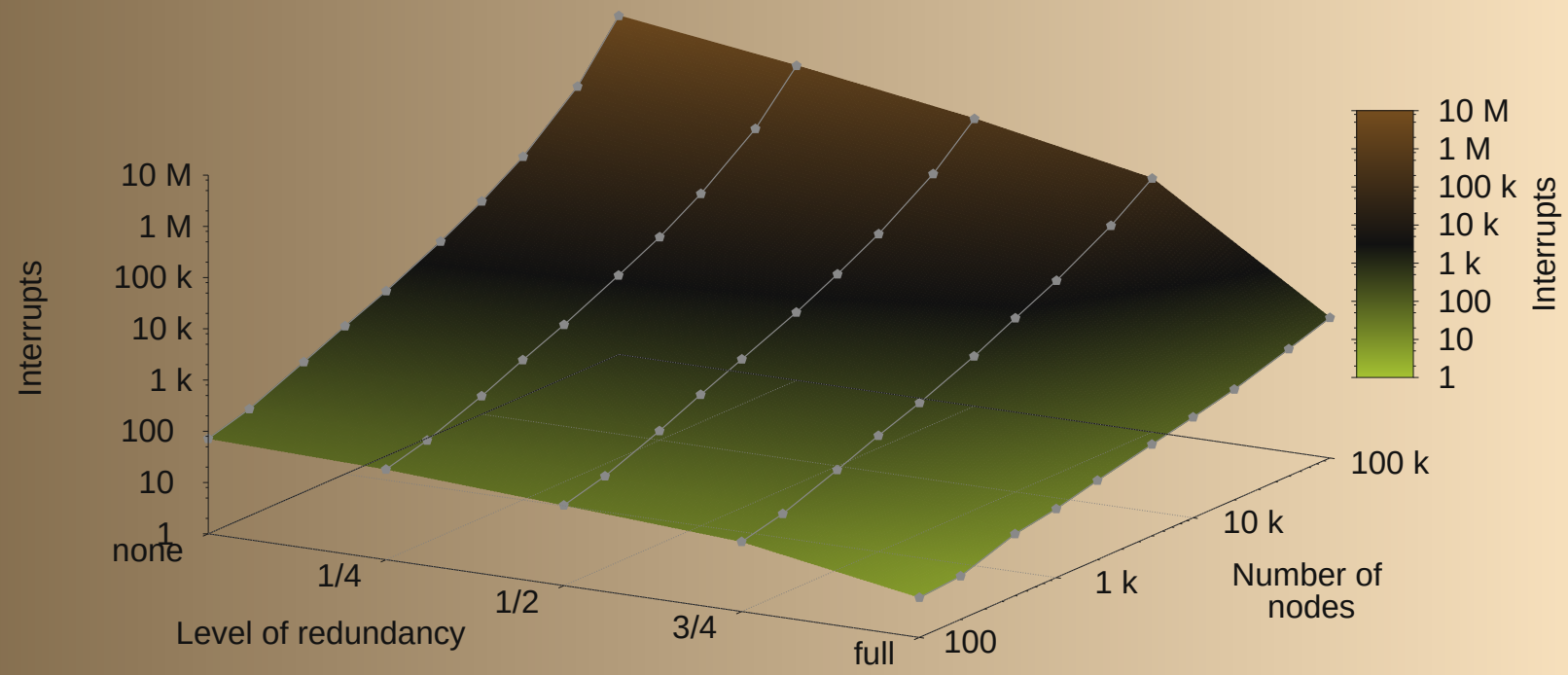
Implications and Trade-Offs

Motivation
Design
Evaluation
Analysis
Implications
Summary

$$R = \frac{T_w + T_{o(\text{none})}}{T_w + T_{o(\text{full})} + T_{r\text{MPI}}} \quad (1)$$

- T_w: amount of work
- $T_{o(\text{none})}$: checkpoint, restart, rework overhead without redundant nodes
- $T_{o(\text{full})}$: overhead with redundant nodes
- If $R > 2$ then using redundant nodes makes sense

Motivation
Design
Evaluation
Analysis
Implications
Summary



5,000h work, 5 min checkpoint, 10 min restart, 1-year node MTBF

Summary

Summary

Motivation
Design
Evaluation
Analysis
Implications
Summary

- Redundant MPI library can be done at user level
- Overhead for applications is not significant for most applications
- Application restart simulator allows modeling of
  - varying node counts
  - node MTBF
  - level of redundancy
  - failure distribution function (exponential, gamma, weibull)
  - amount of work, checkpoint and restart times
- Model helps determine when redundancy pays off