ACR: AUTOMATIC CHECKPOINT/RESTART FOR SOFT AND HARD ERROR PROTECTION.

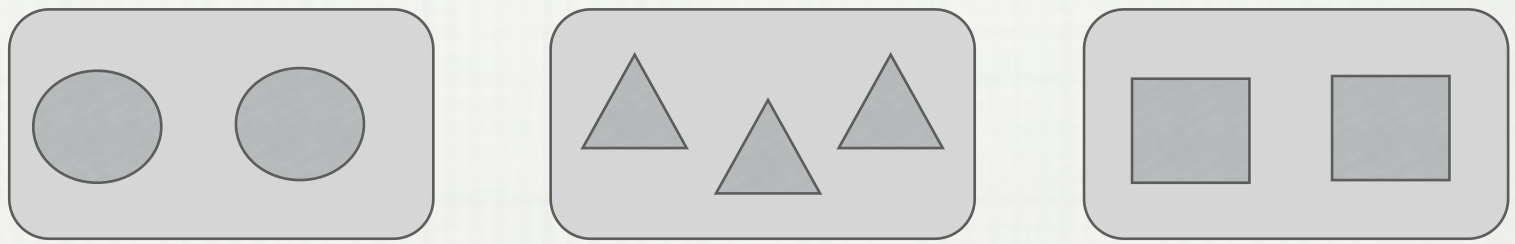
Xiang Ni, Esteban Meneses, Nikhil Jain, Sanjay Kale

Parallel Programming Lab, UIUC

CONTENTS

BACKGROUND

TASK

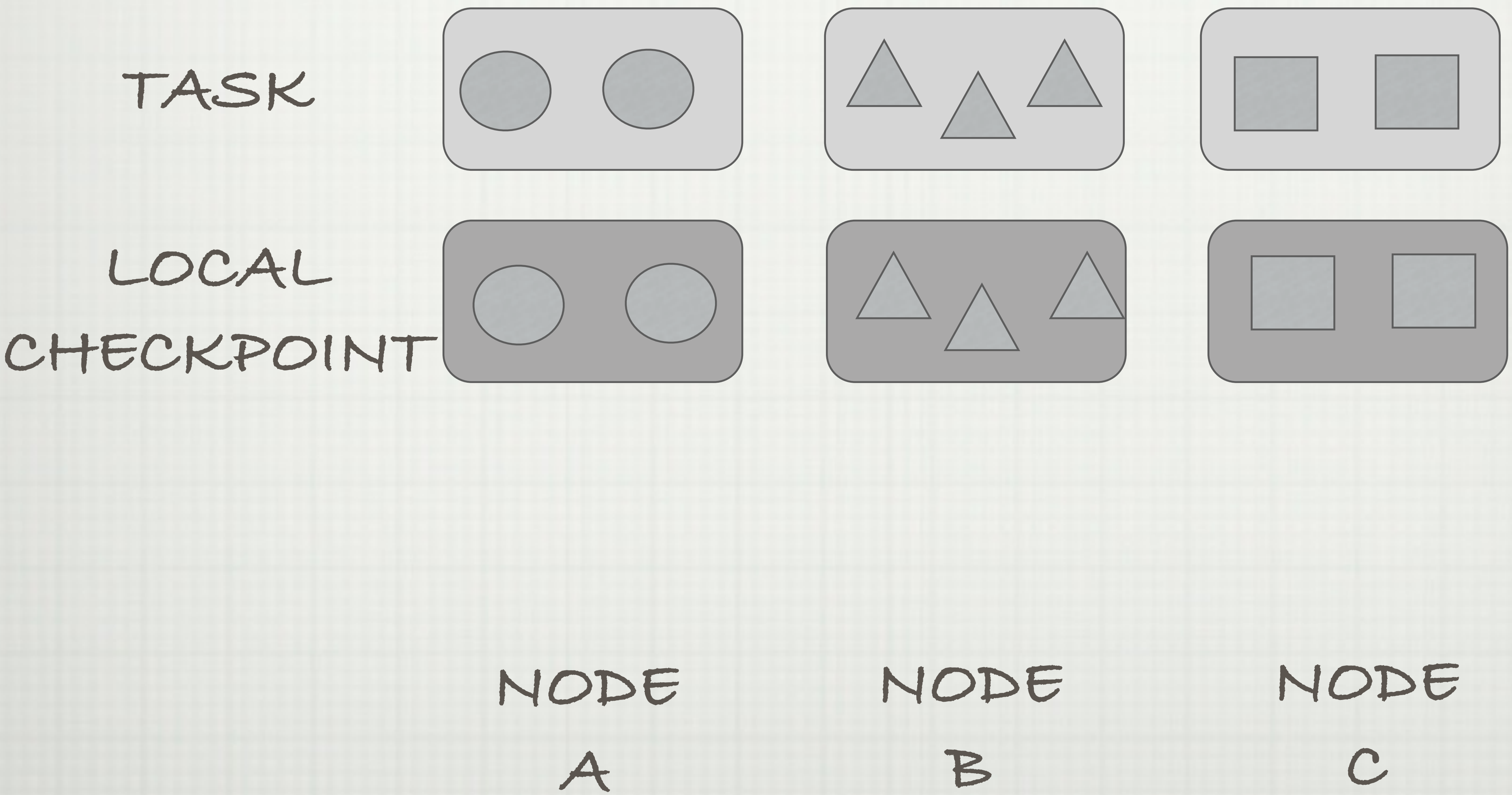


NODE A

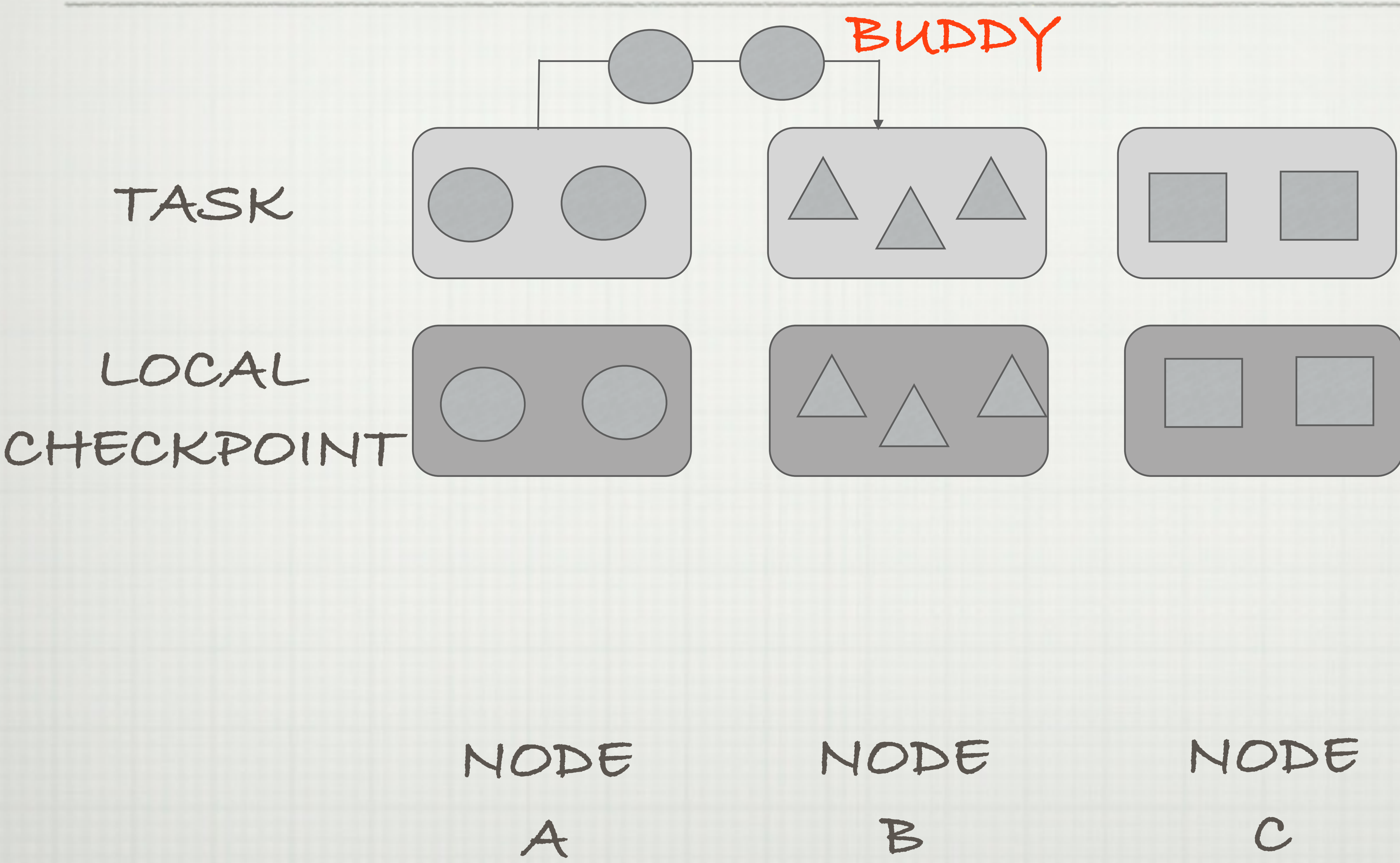
NODE B

NODE C

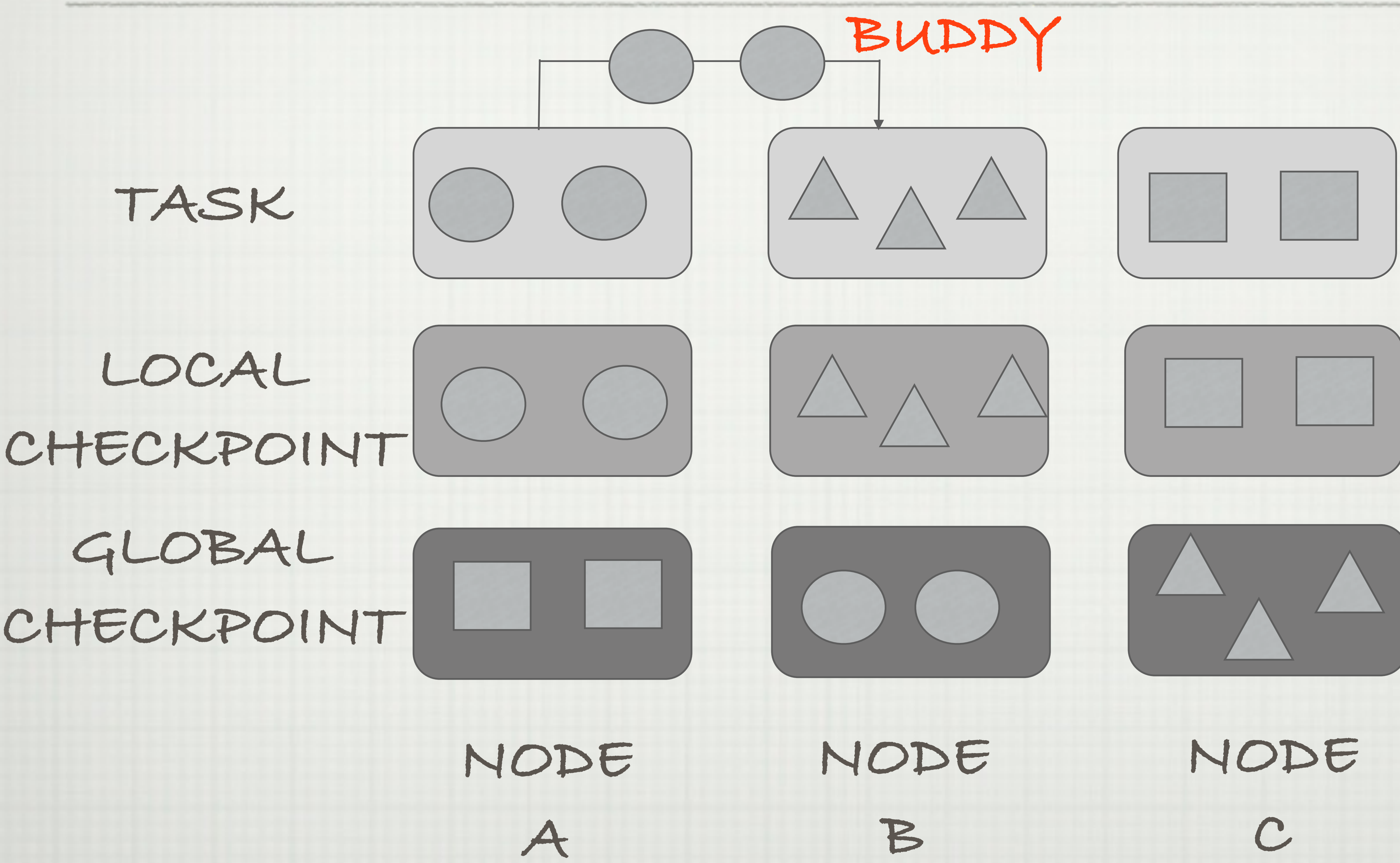
BACKGROUND



BACKGROUND



BACKGROUND


BUDDY
TASK
LOCAL CHECKPOINT
GLOBAL CHECKPOINT
NODE A
NODE B
NODE C

MOTIVATION

- New challenge: soft error.
  - An unintended change in the state of an electronic device that alters the information that it stores without destroying its functionality.
  - Detectable and correctable: single bit flip
  - Detectable and uncorrectable: double bit flips
  - Undetectable, uncorrectable and incorrect program outcome

MOTIVATION

- New challenge: soft error.
  - An unintended change in the state of an electronic device that alters the information that it stores without destroying its functionality.
  - Detectable and correctable: single bit flip
  - Detectable and uncorrectable: double bit flips
  - Undetectable, uncorrectable and incorrect

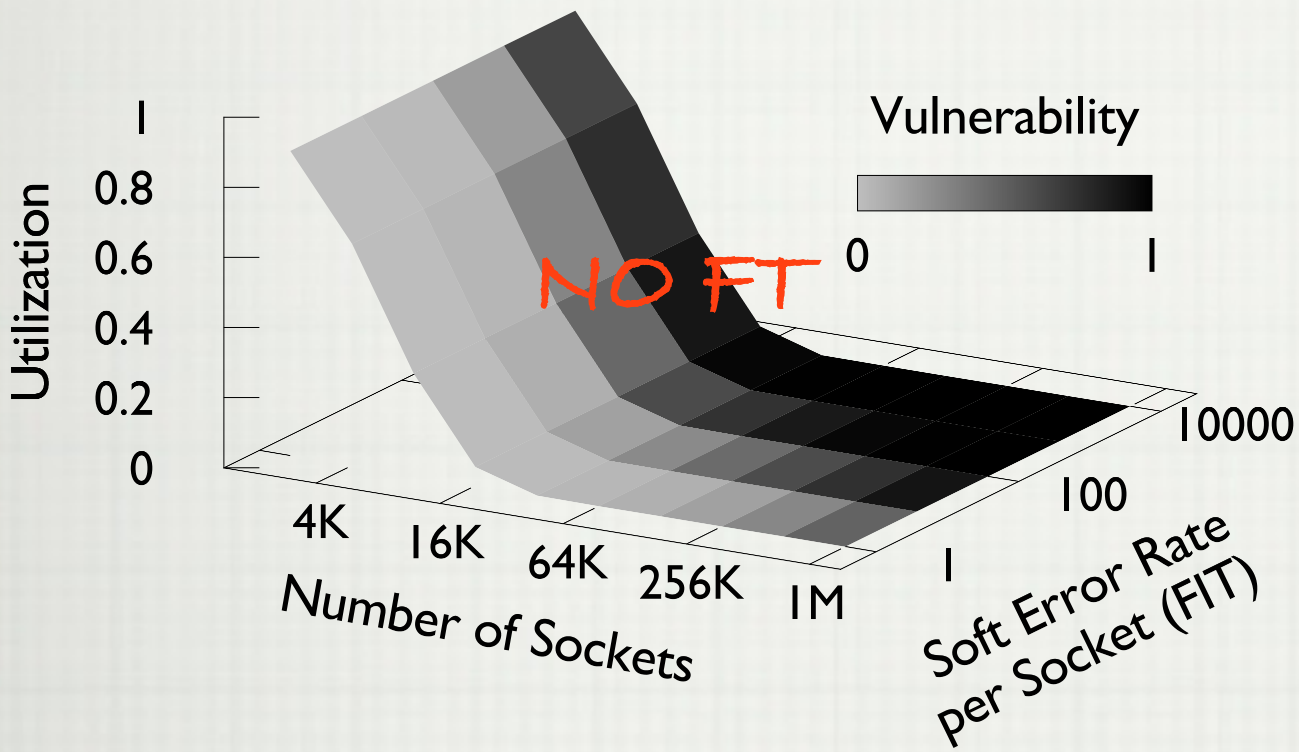
SILENT DATA CORRUPTION

MOTIVATION

- Why soft failure rate will increase?
  - Computer electronic's sensitivity to radiation increases as their dimensions and operating voltage decreases
    - The requirements for high performance and low power.
- What may happen if soft failure rate keeps increasing?

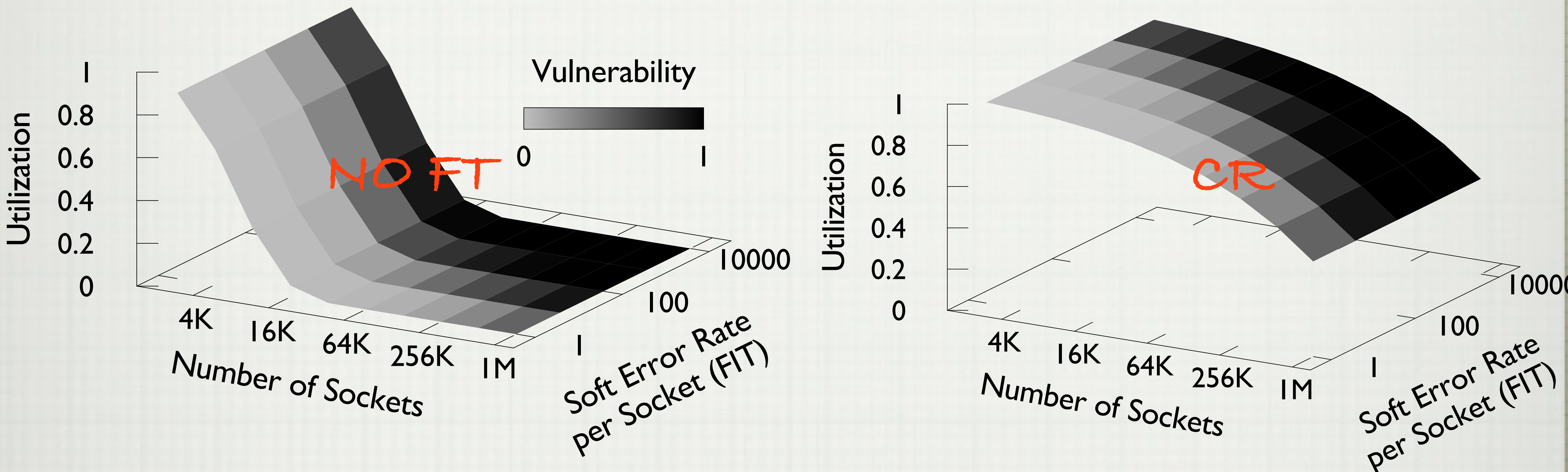
MOTIVATION

MOTIVATION



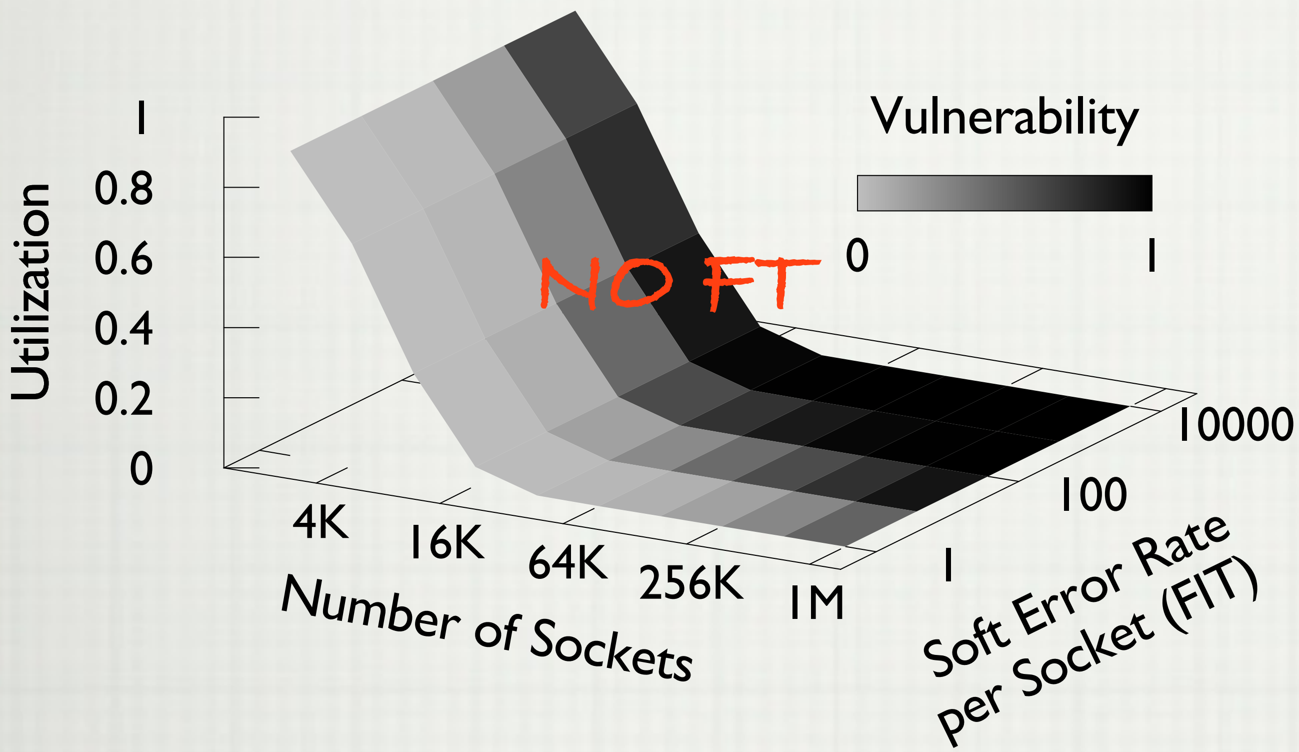
Vulnerability
0
1
NO FT
Utilization
1
0.8
0.6
0.4
0.2
0
4K
16K
64K
256K
1M
Number of Sockets
1
100
10000
Soft Error Rate per Socket (FIT)

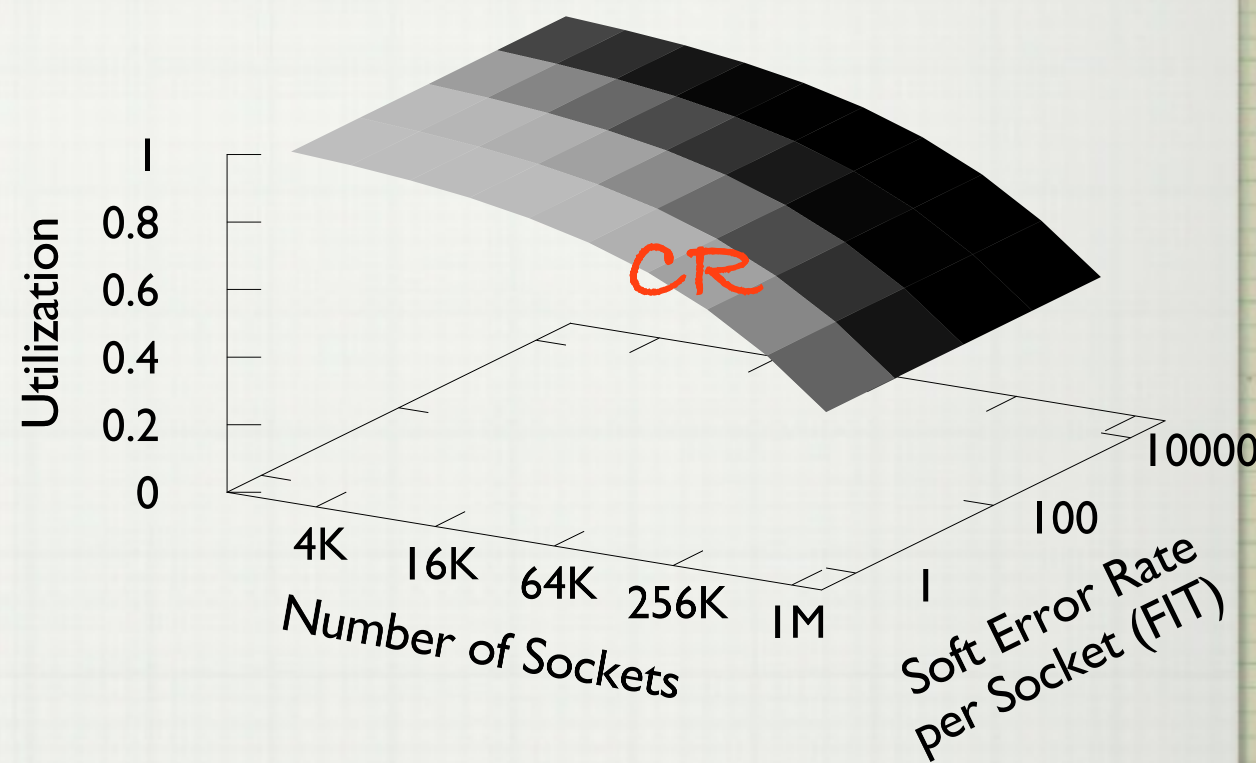
MOTIVATION

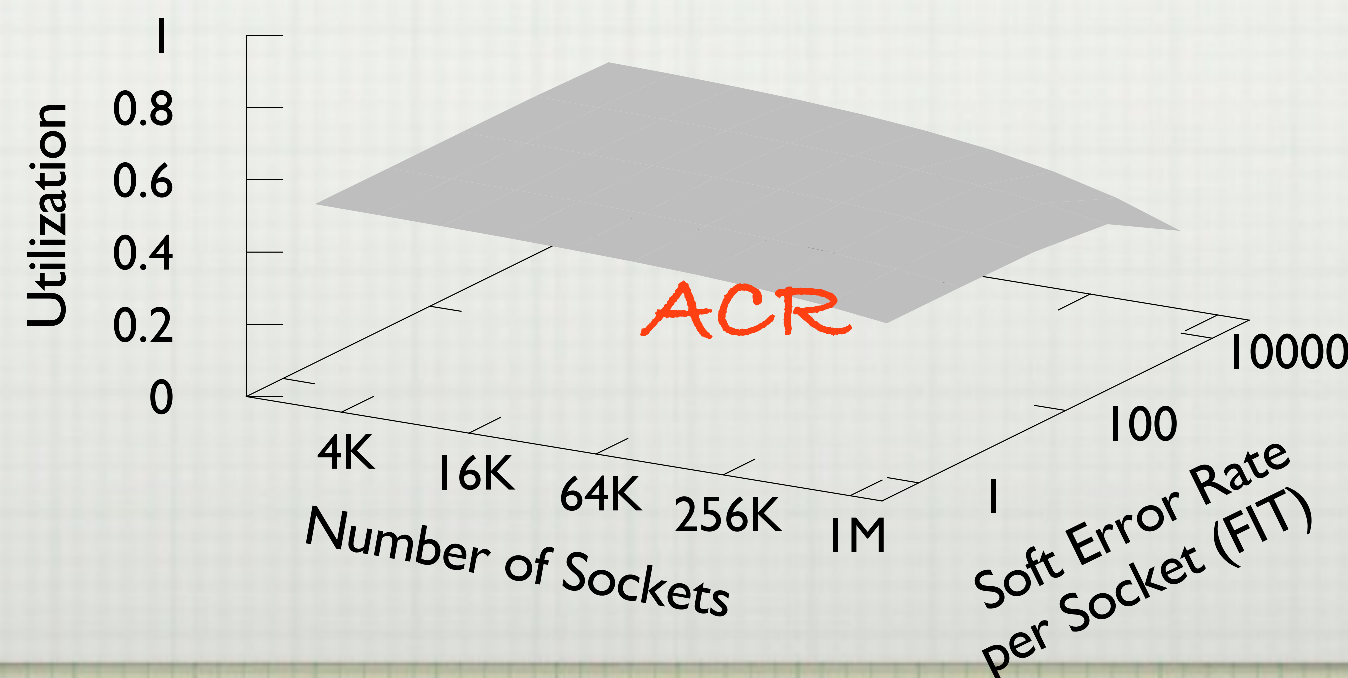


Vulnerability
0
1
NO FT
Utilization
0
0.2
0.4
0.6
0.8
1
4K
16K
64K
256K
1M
Number of Sockets
1
100
10000
Soft Error Rate per Socket (FIT)
CR
Utilization
0
0.2
0.4
0.6
0.8
1
4K
16K
64K
256K
1M
Number of Sockets
1
100
10000
Soft Error Rate per Socket (FIT)

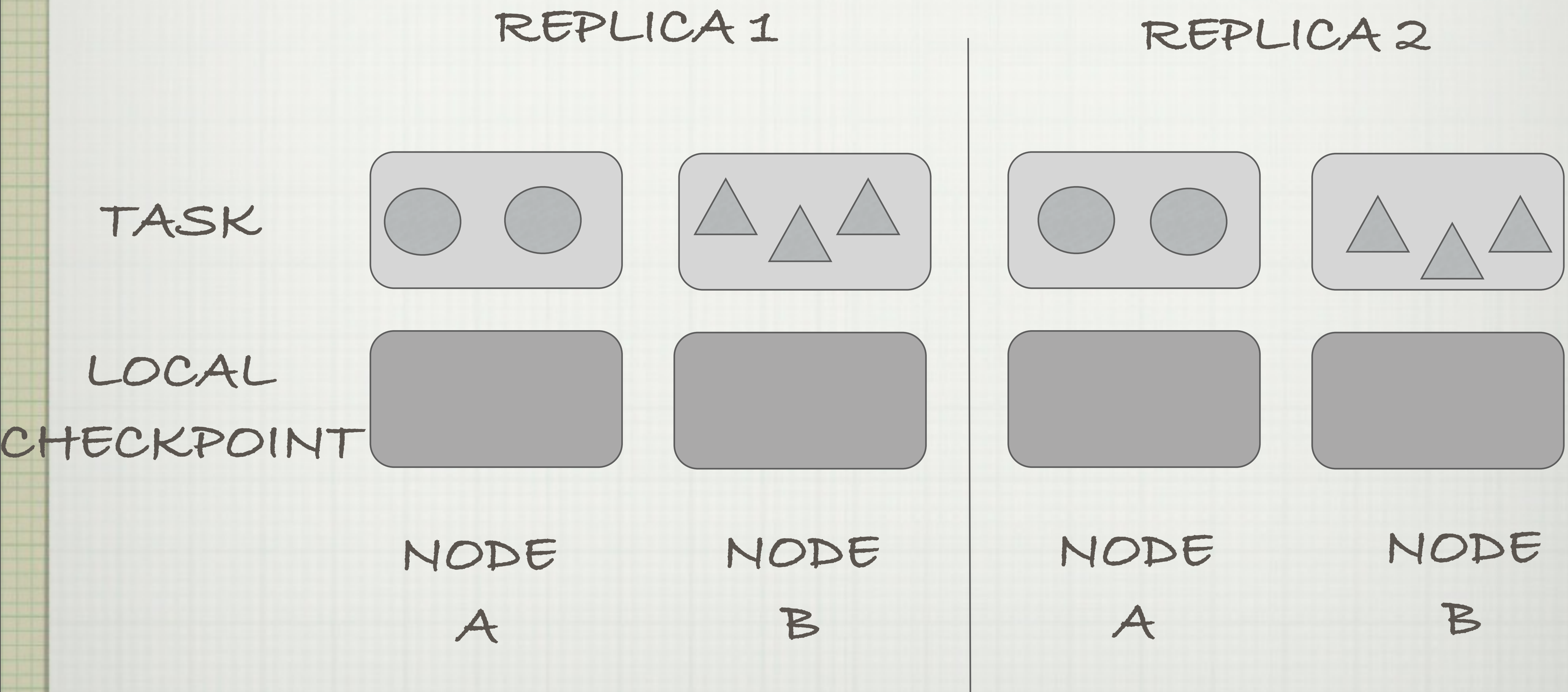
MOTIVATION


Vulnerability
0
1
NO FT
Utilization
0
0.2
0.4
0.6
0.8
1
4K
16K
64K
256K
1M
Number of Sockets
1
100
10000
Soft Error Rate per Socket (FIT)

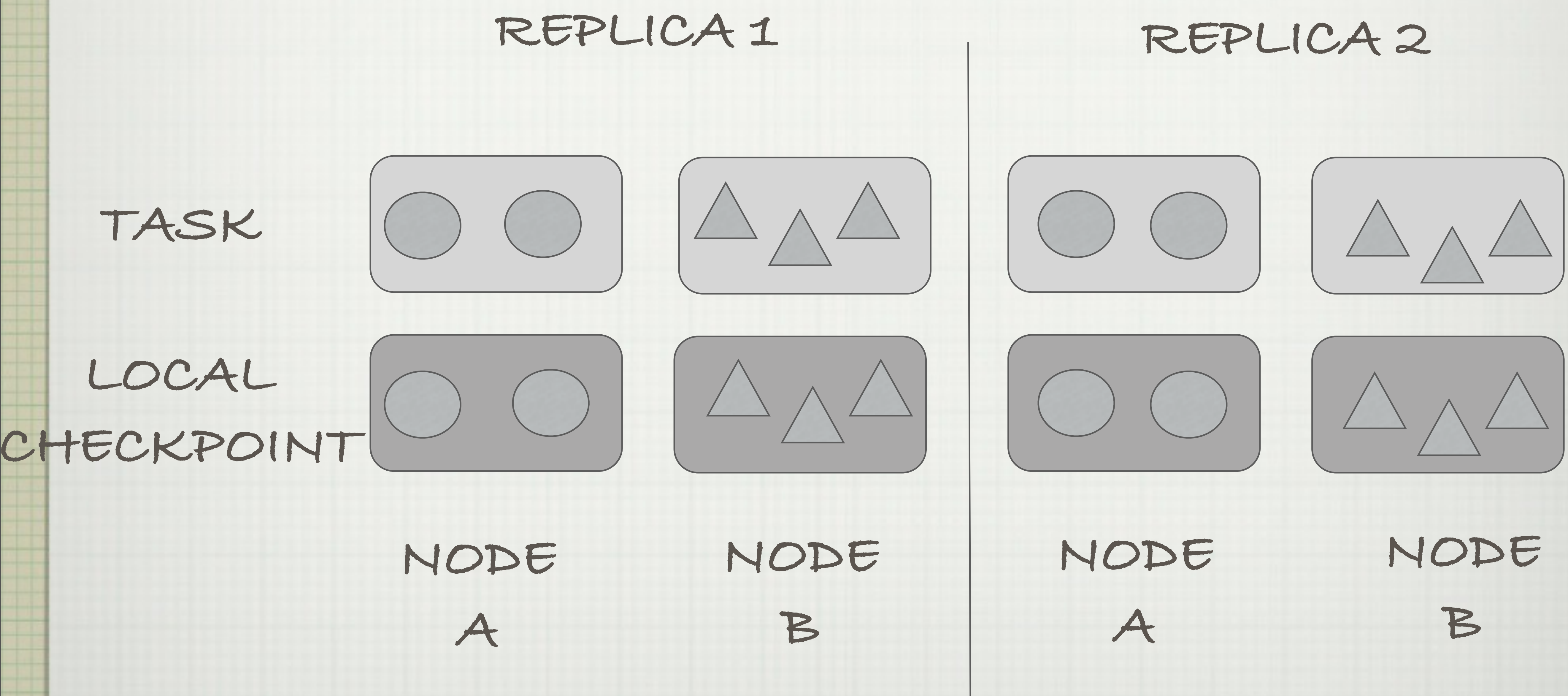

CR
Utilization
0
0.2
0.4
0.6
0.8
1
4K
16K
64K
256K
1M
Number of Sockets
1
100
10000
Soft Error Rate per Socket (FIT)


ACR
Utilization
0
0.2
0.4
0.6
0.8
1
4K
16K
64K
256K
1M
Number of Sockets
1
100
10000
Soft Error Rate per Socket (FIT)

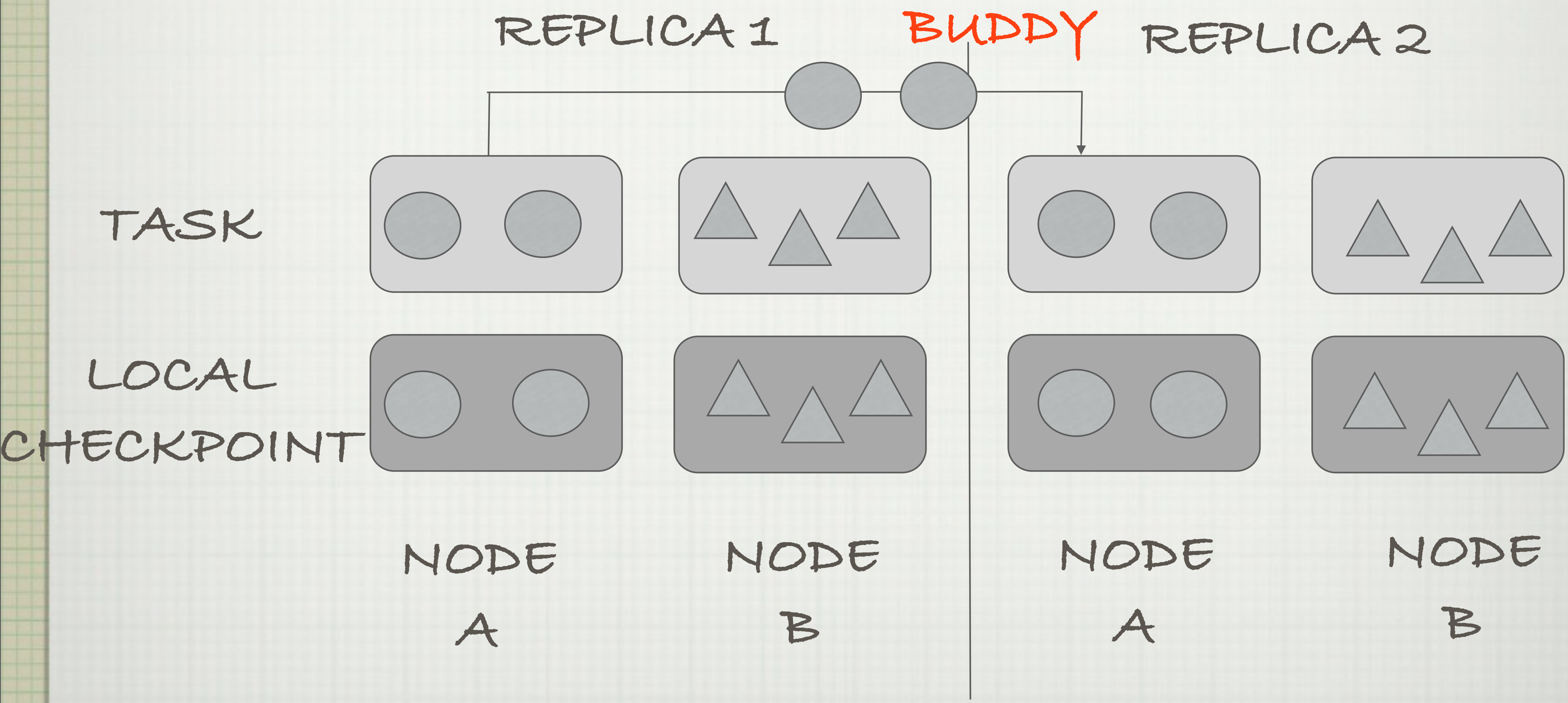
REPLICATION ENHANCED CHECKPOINTING



REPLICATION ENHANCED CHECKPOINTING

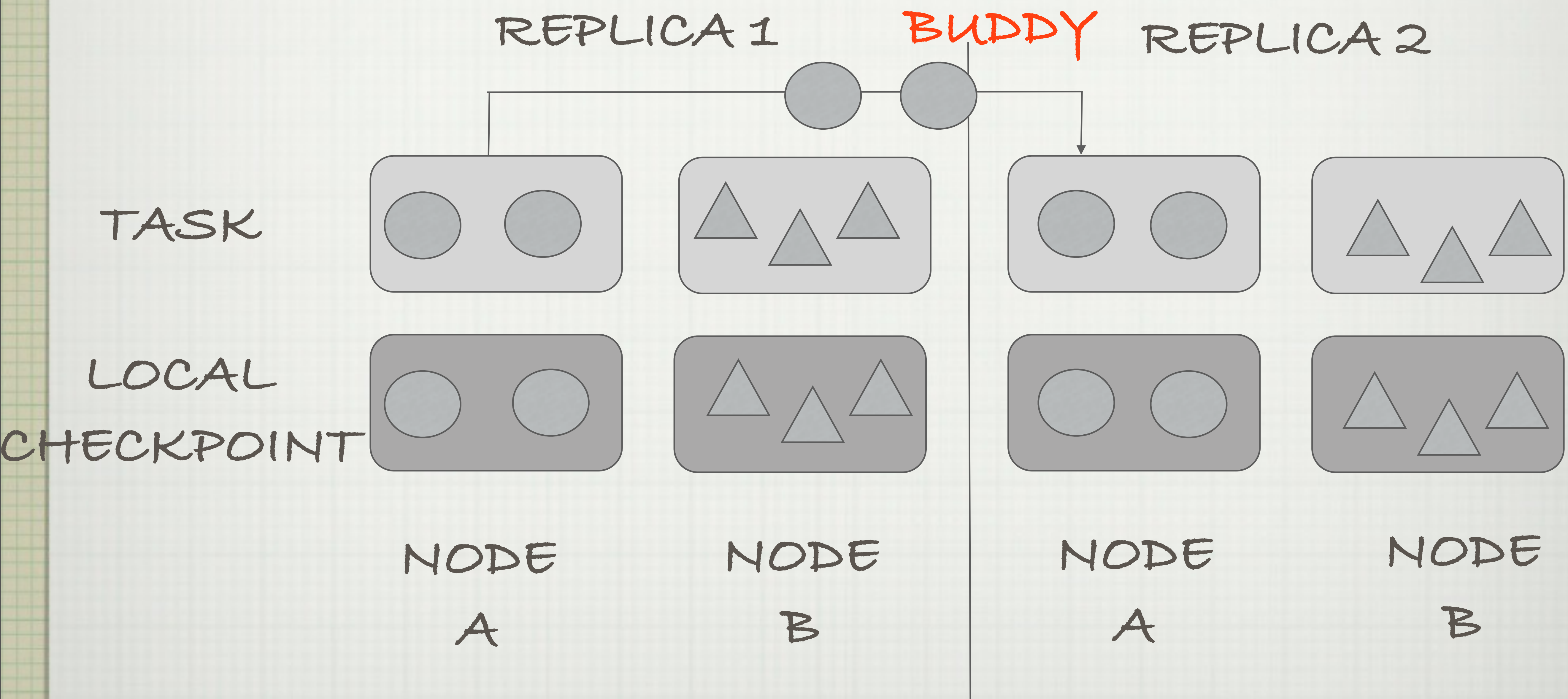


REPLICATION ENHANCED CHECKPOINTING



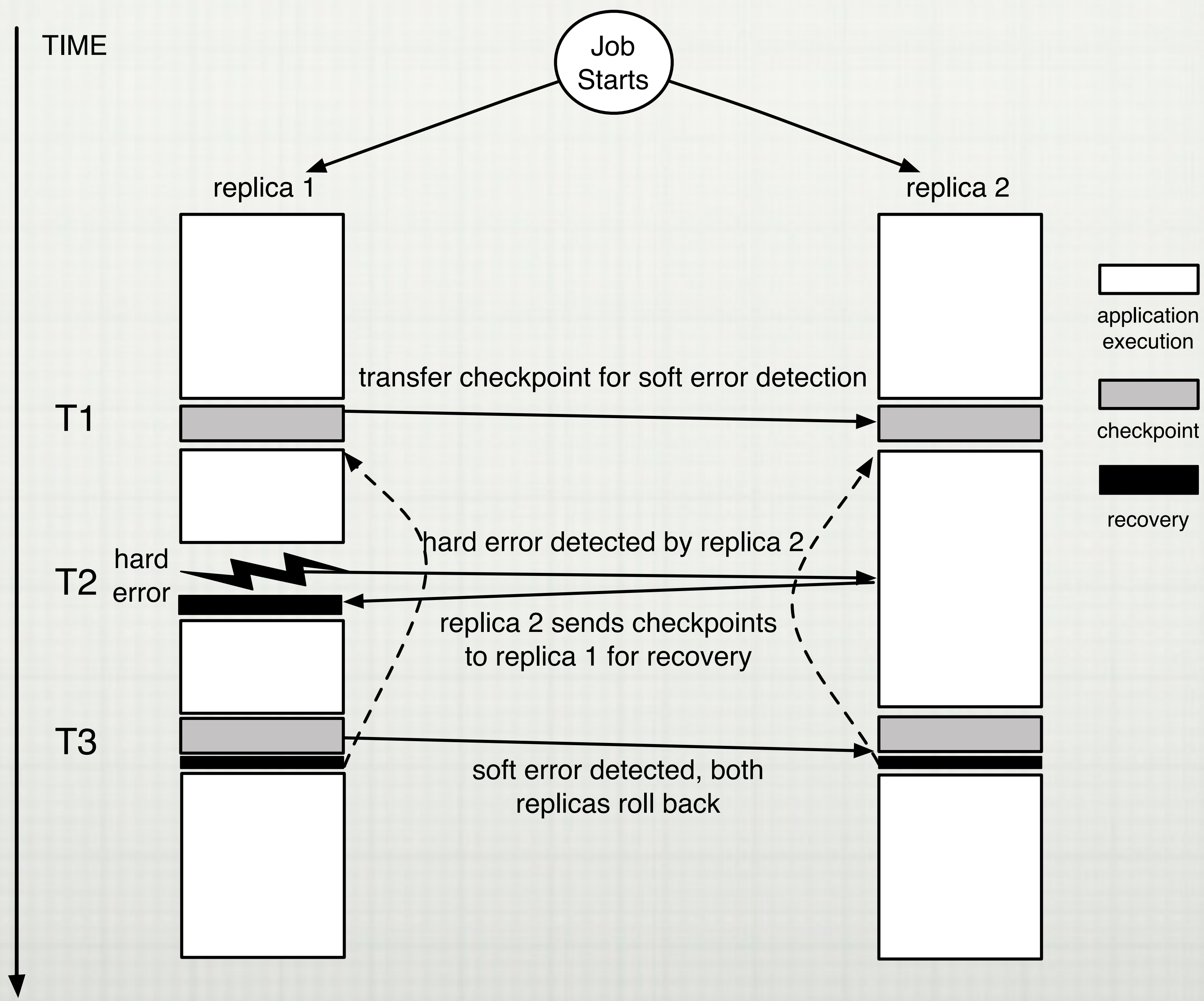
7

REPLICATION ENHANCED CHECKPOINTING



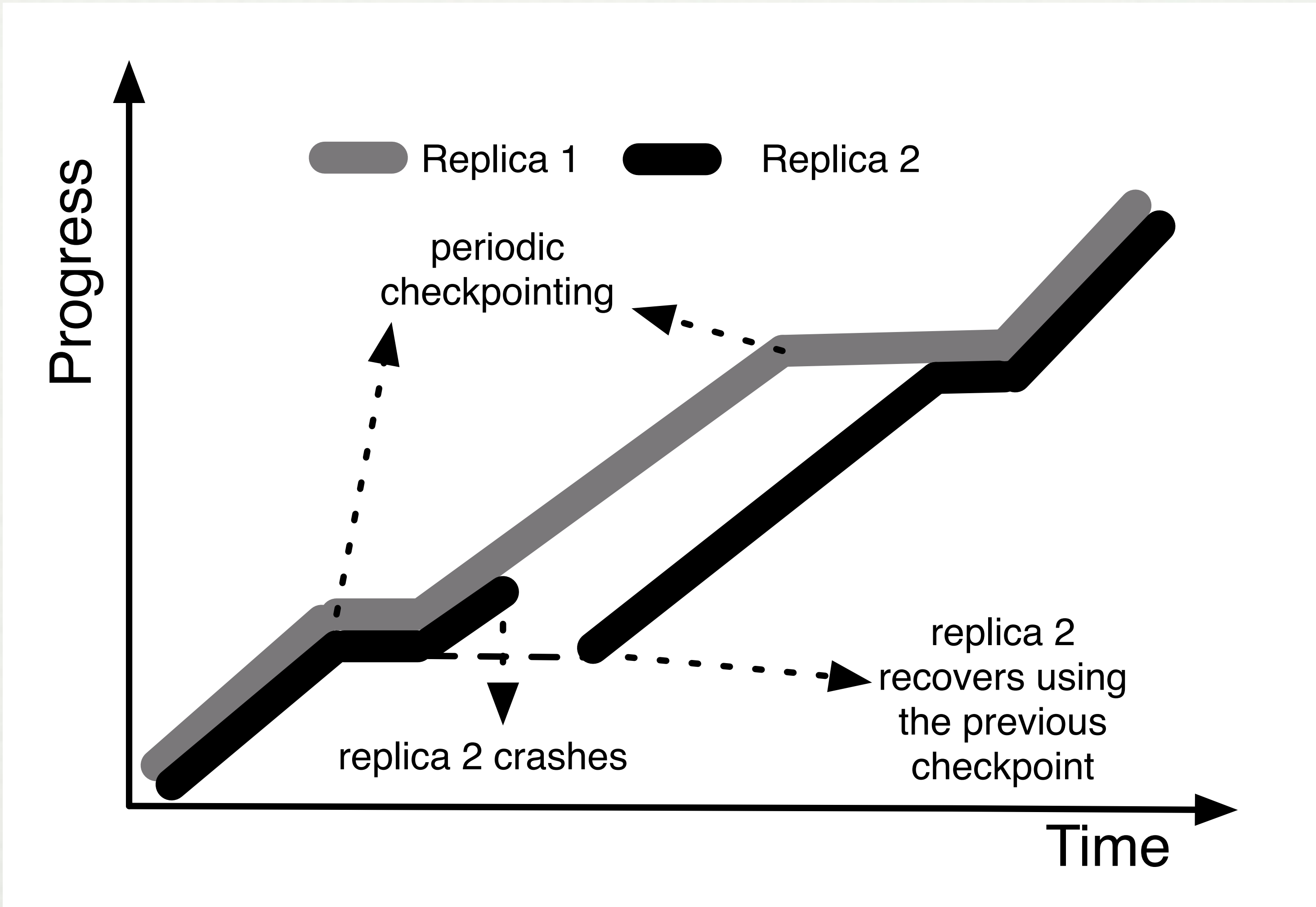
Comparison for soft data corruption

REPLICATION ENHANCED CHECKPOINTING



8

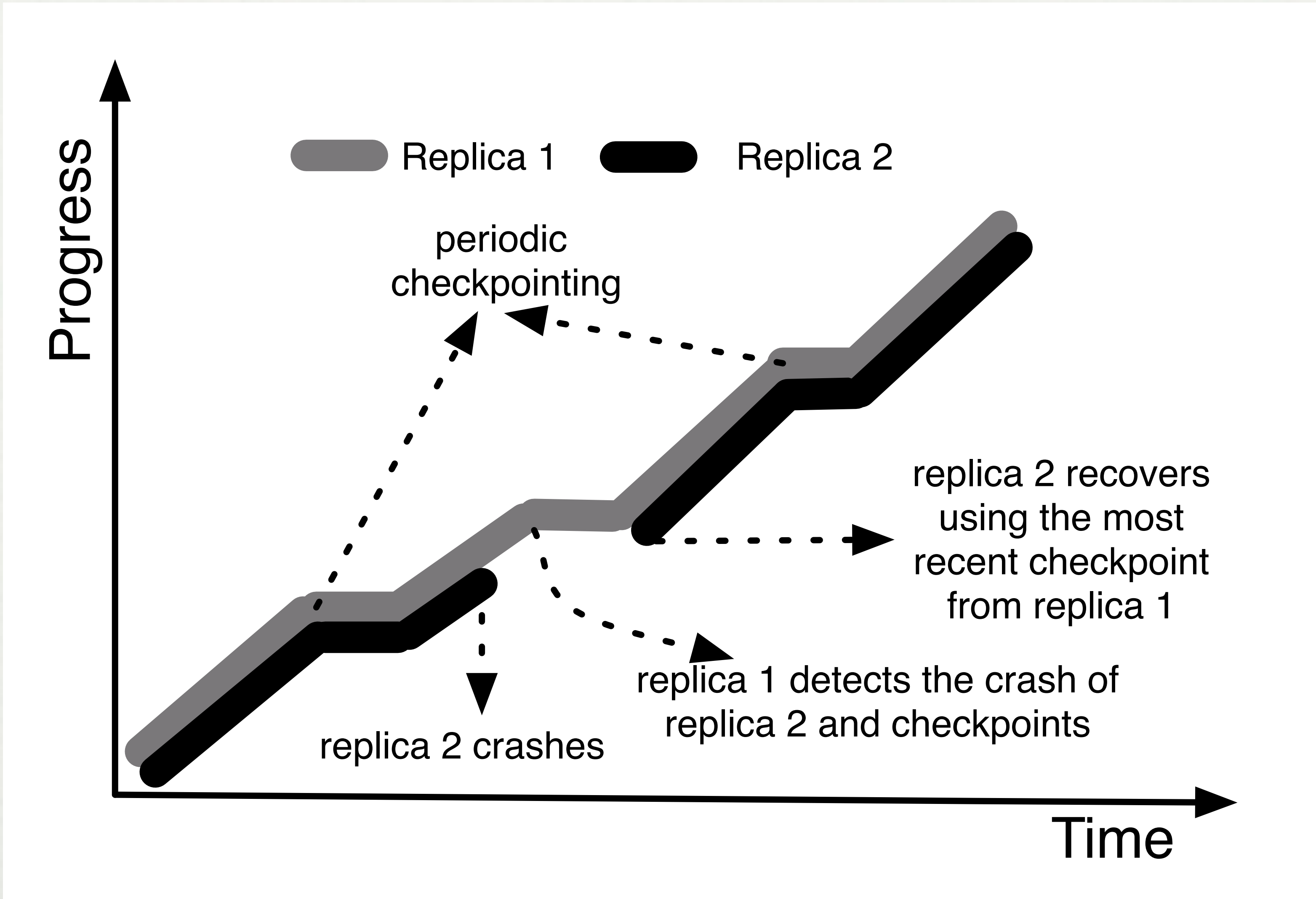
DIFFERENT WAYS TO RESTART FROM HARD ERROR



Strong resilience

Recover from the previous checkpoint

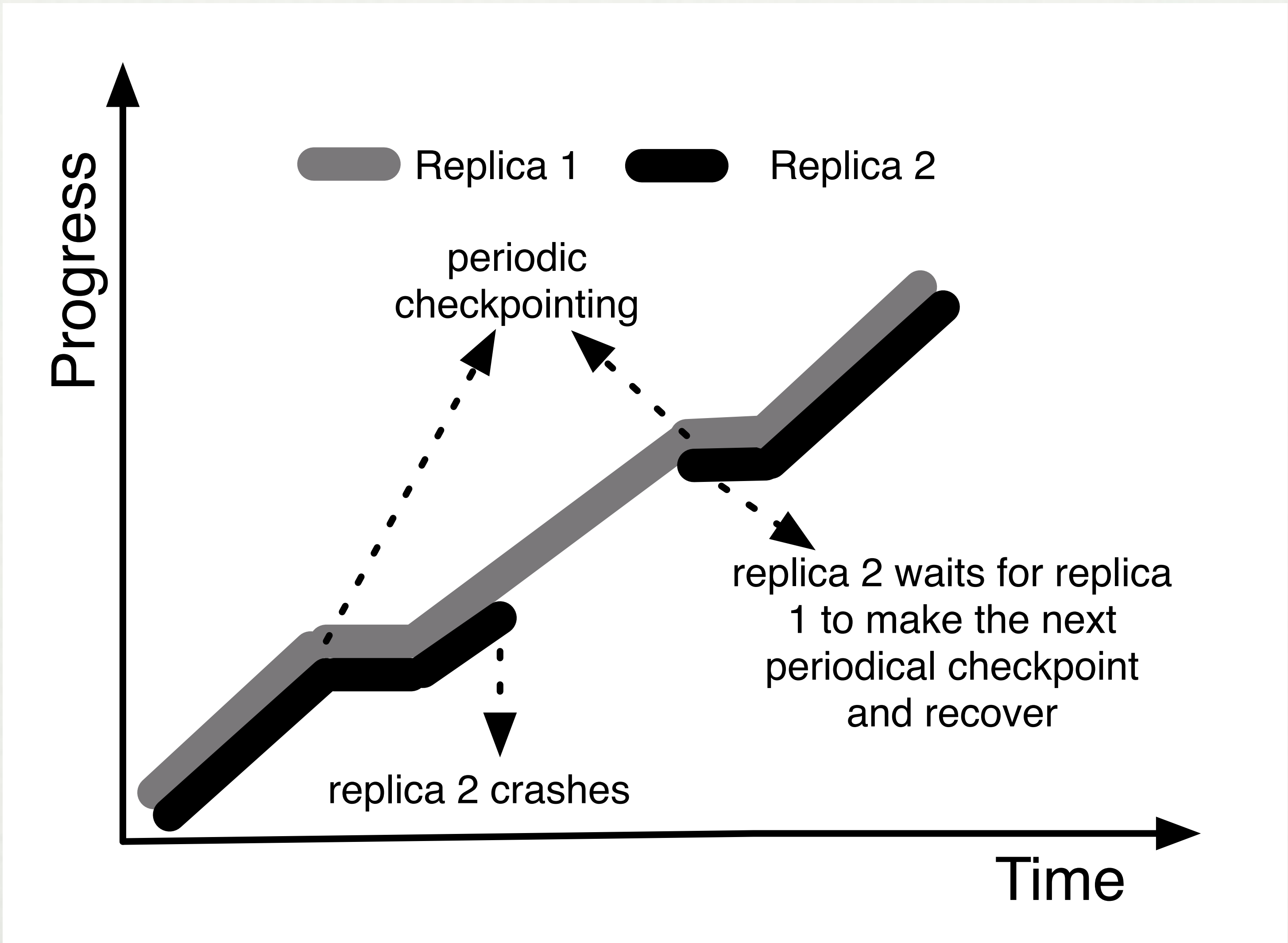
DIFFERENT WAYS TO RESTART FROM HARD ERROR



Medium resilience

Healthy replica schedules an immediate checkpoint

DIFFERENT WAYS TO RESTART FROM HARD ERROR



Weak resilience

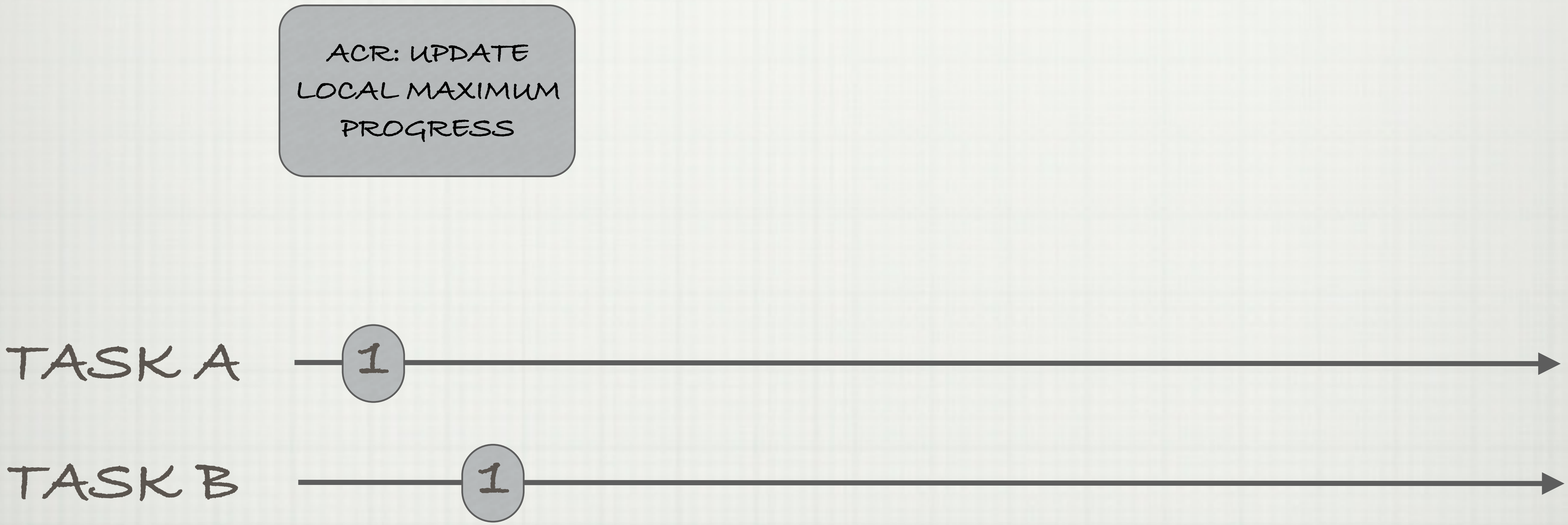
Recover from the next checkpoint

AUTOMATIC CHECKPOINT DECISION

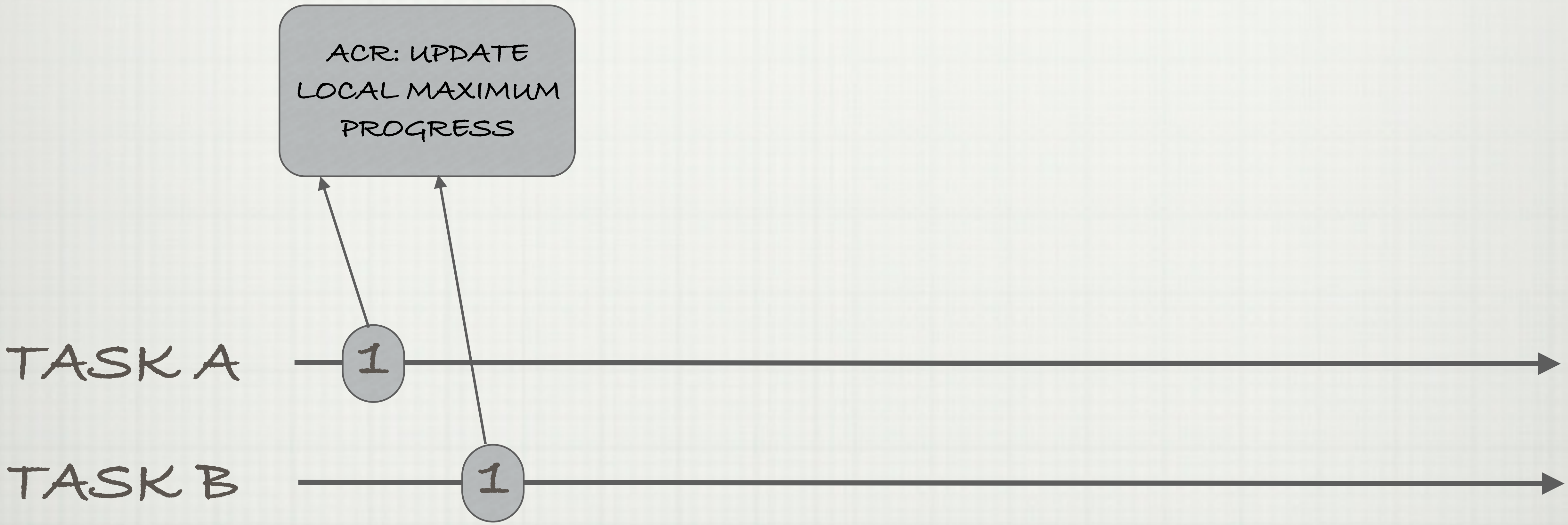
- "Schedule" immediate checkpoint
  - Asynchronous application: no barriers
  - Program will hang after restart if inconsistent states are
- Solution: asynchronous consensus based on scheme available in Charm++*

*MENON ET.AL "AUTOMATED LOAD BALANCING INVOCATION BASED ON APPLICATION CHARACTERISTICS" CLUSTER, 2012

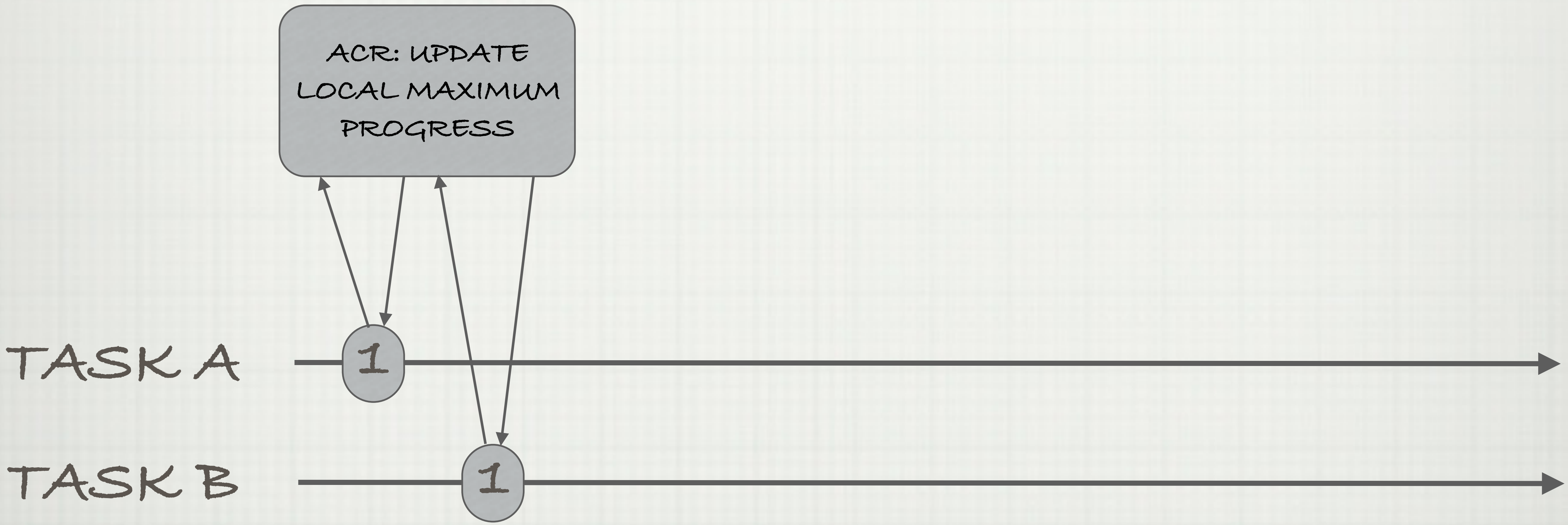
AUTOMATIC CHECKPOINT DECISION



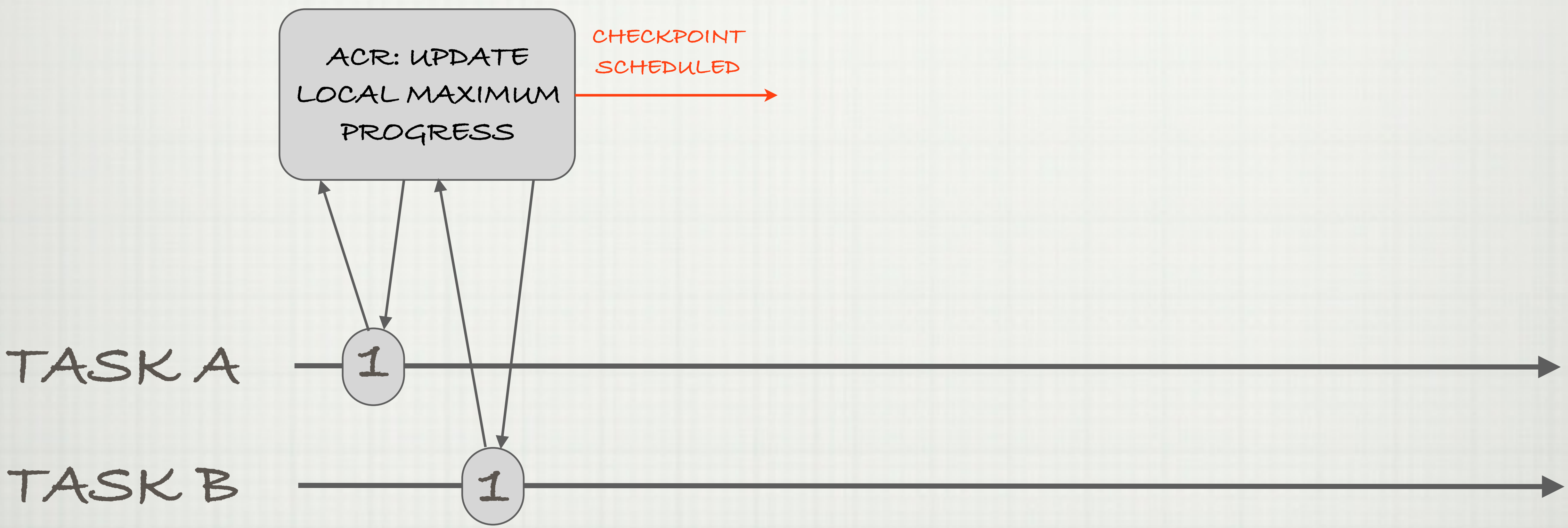
AUTOMATIC CHECKPOINT DECISION



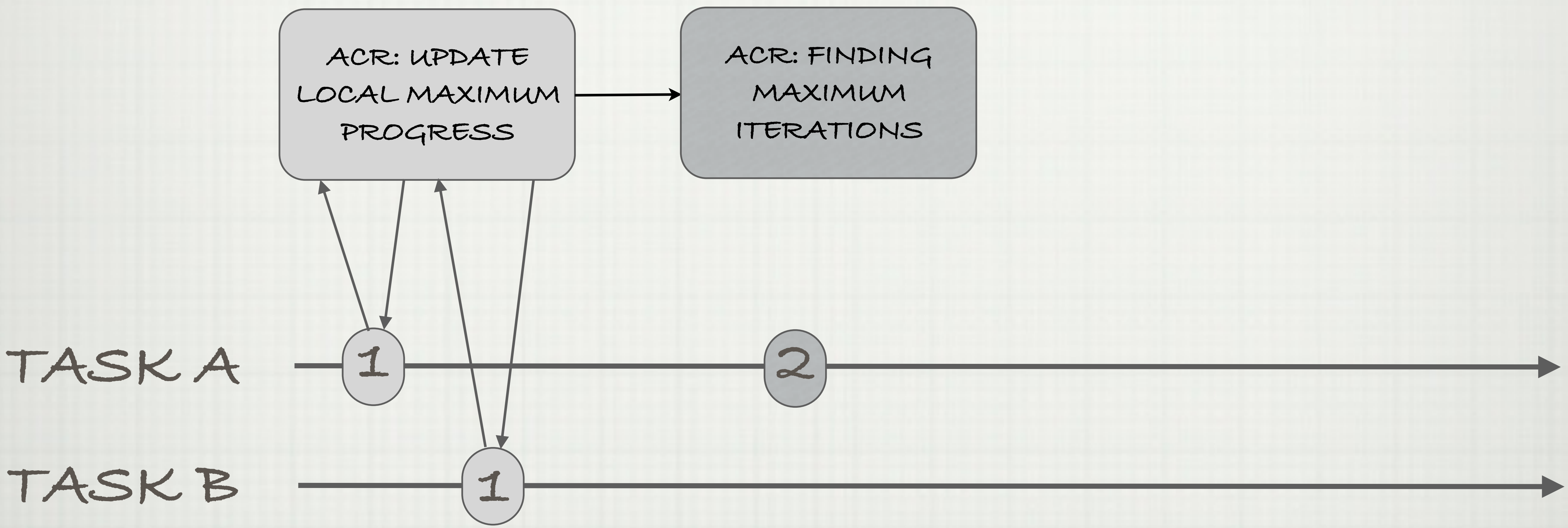
AUTOMATIC CHECKPOINT DECISION



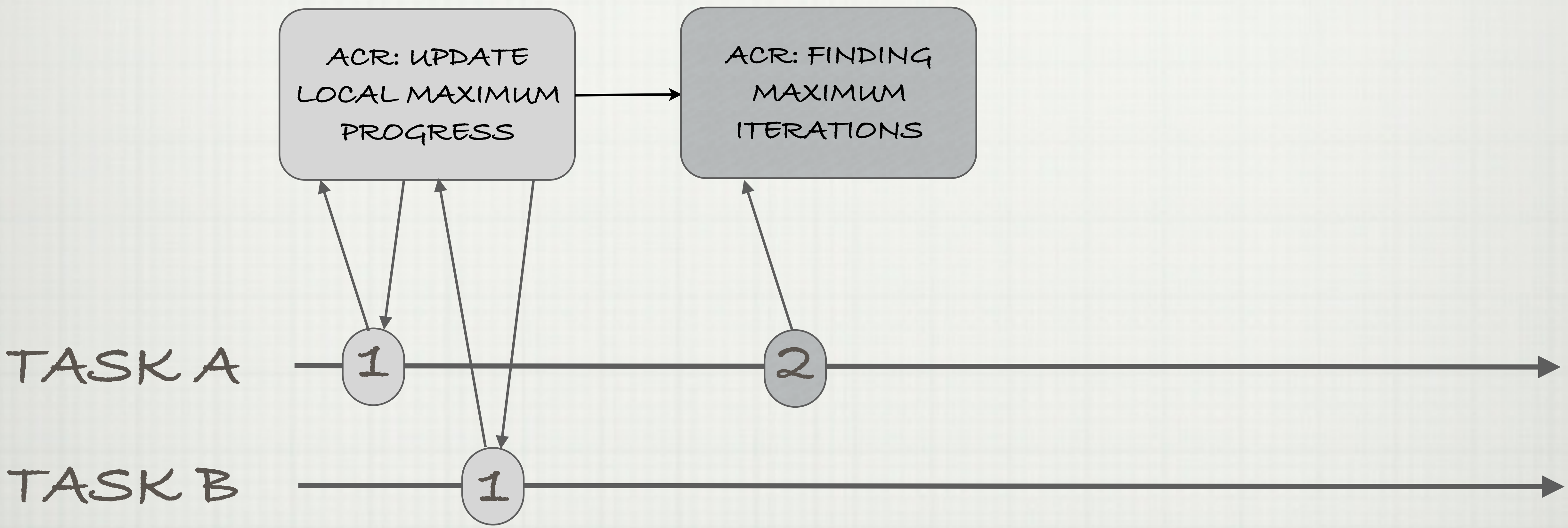
AUTOMATIC CHECKPOINT DECISION



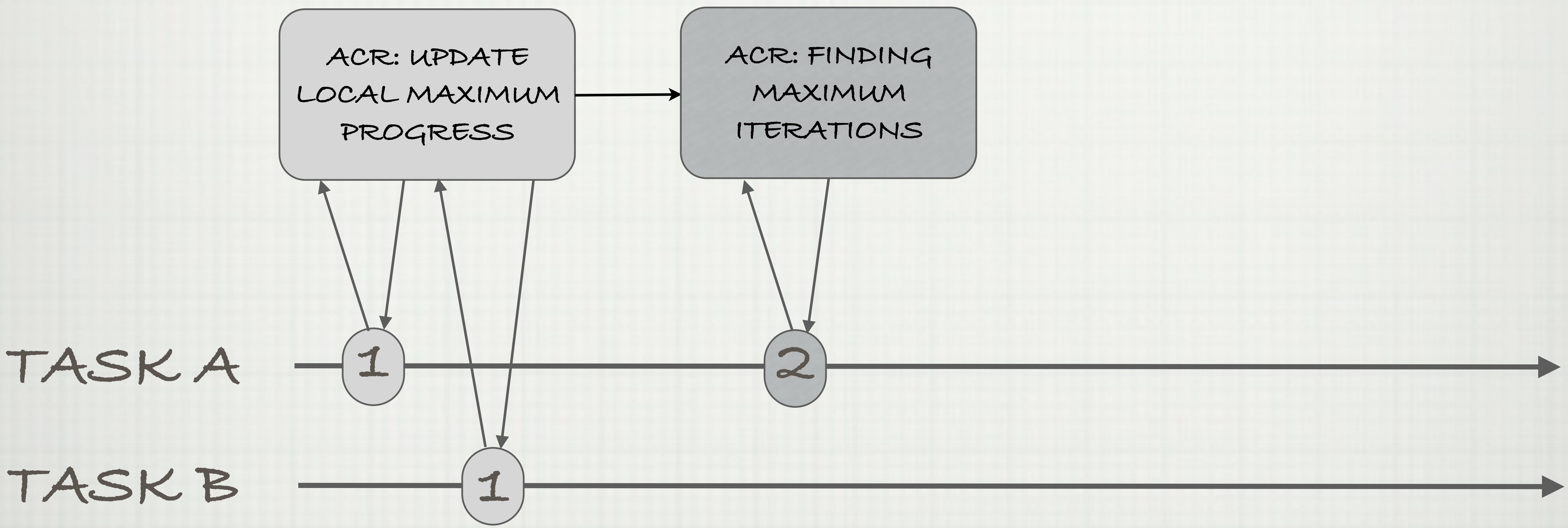
AUTOMATIC CHECKPOINT DECISION



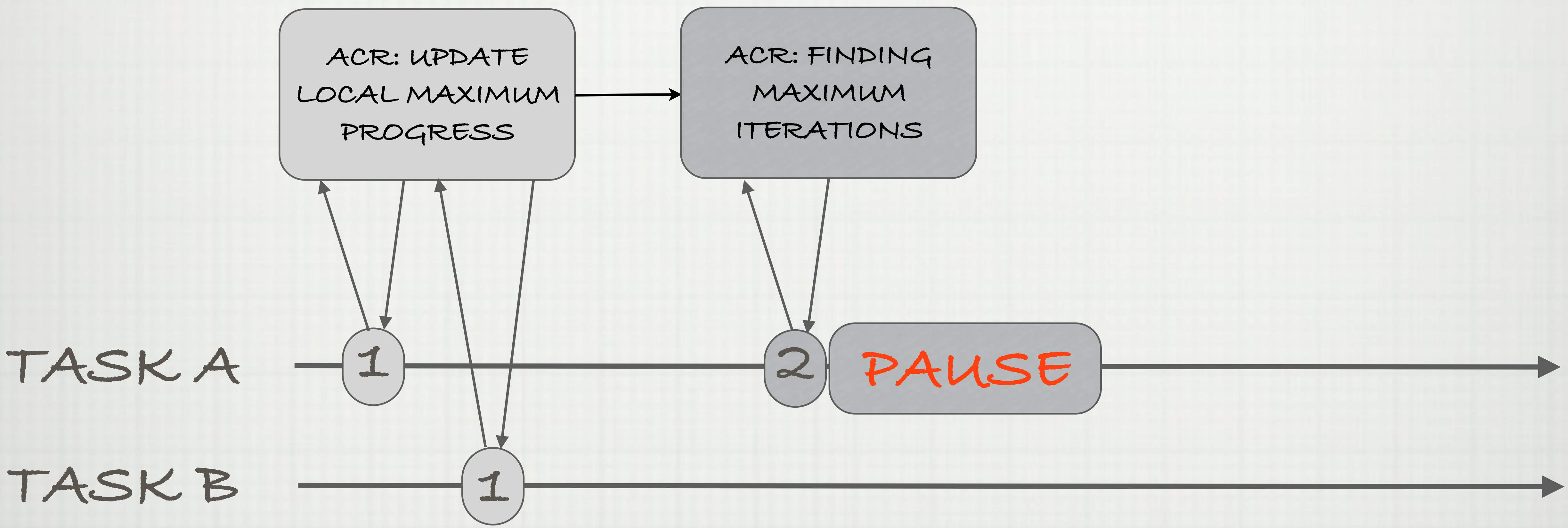
AUTOMATIC CHECKPOINT DECISION



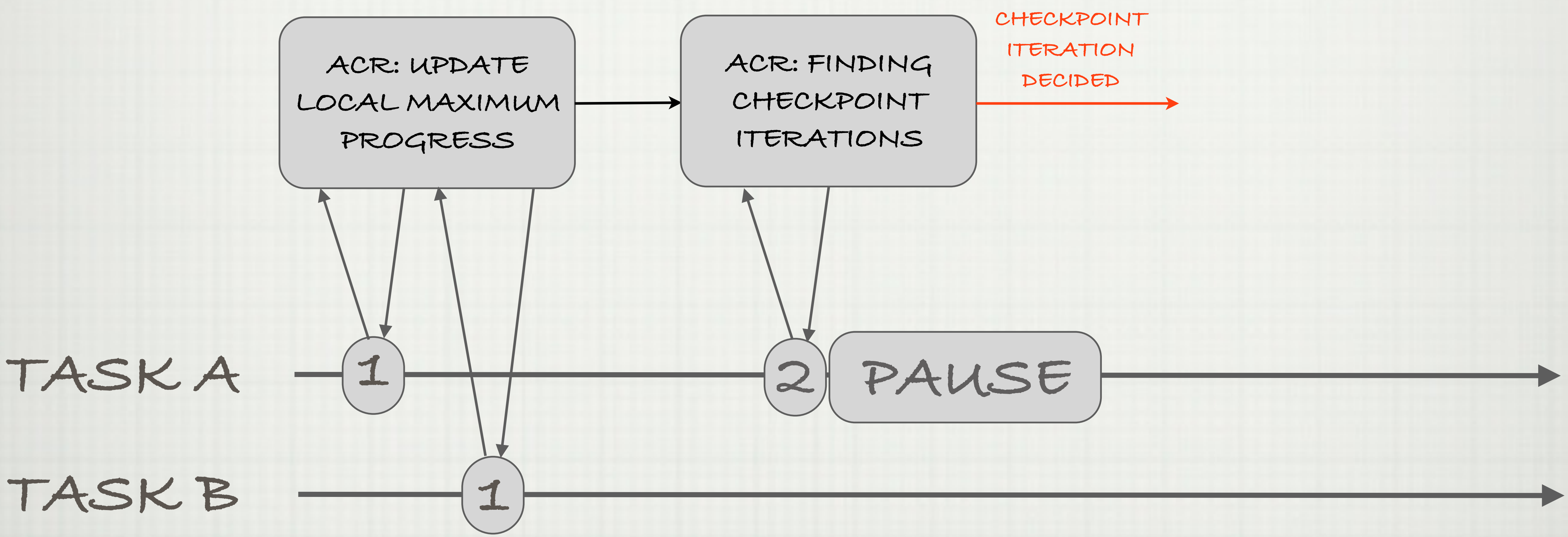
AUTOMATIC CHECKPOINT DECISION



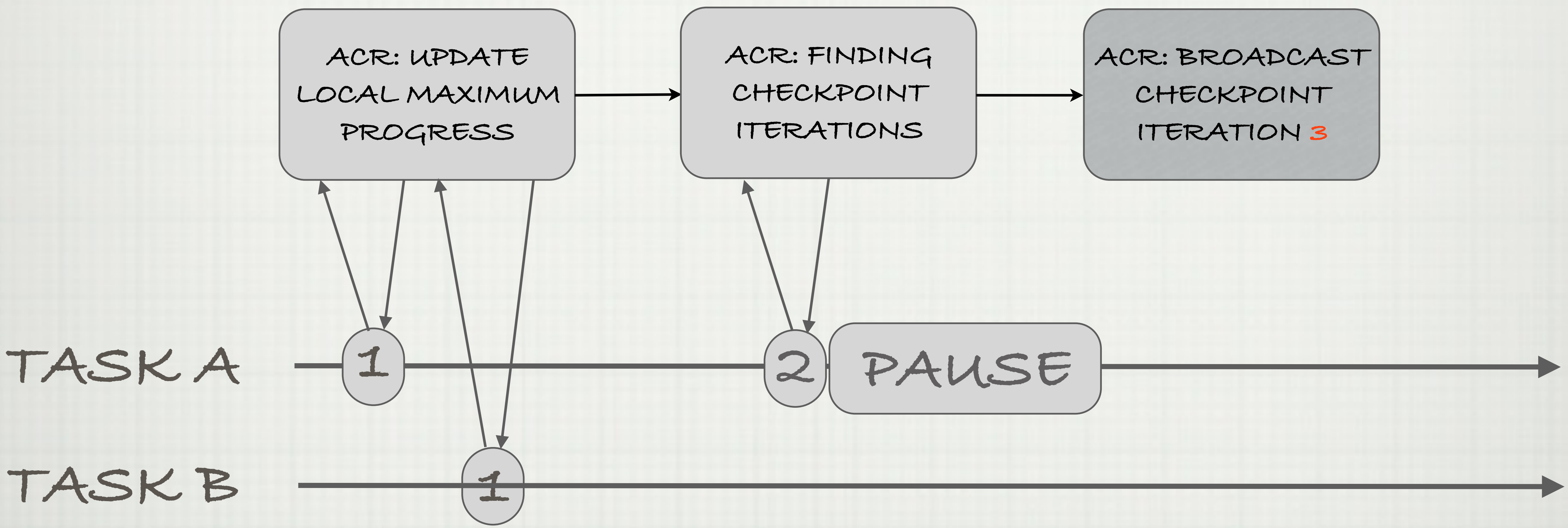
AUTOMATIC CHECKPOINT DECISION



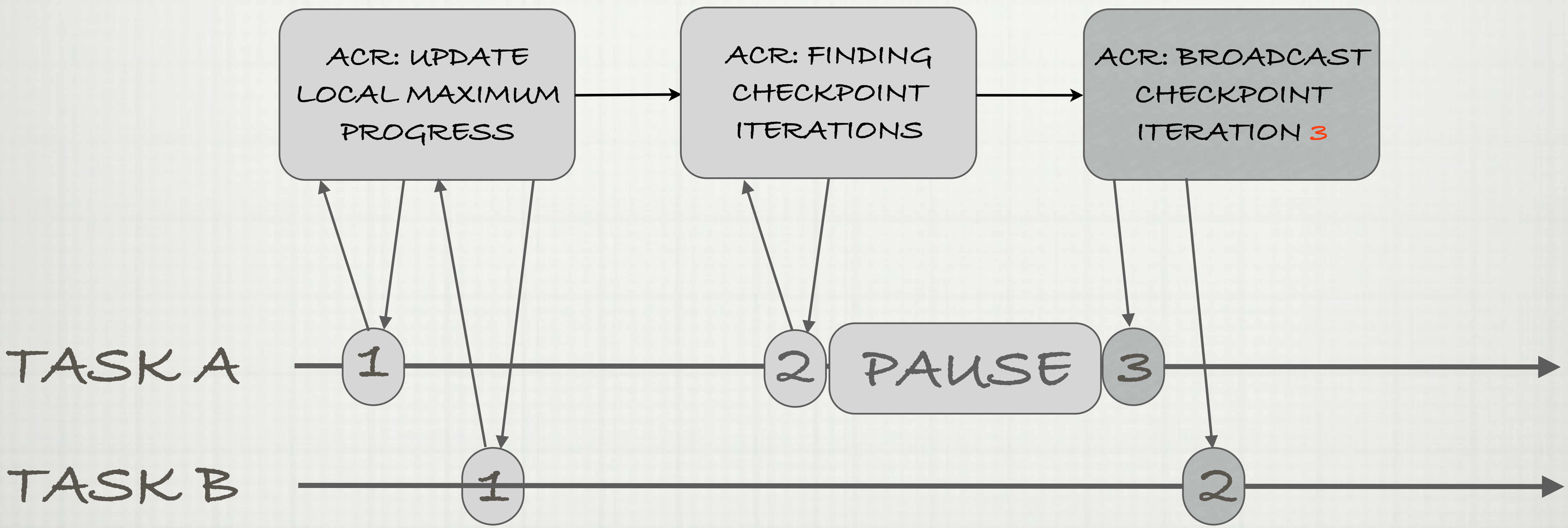
AUTOMATIC CHECKPOINT DECISION



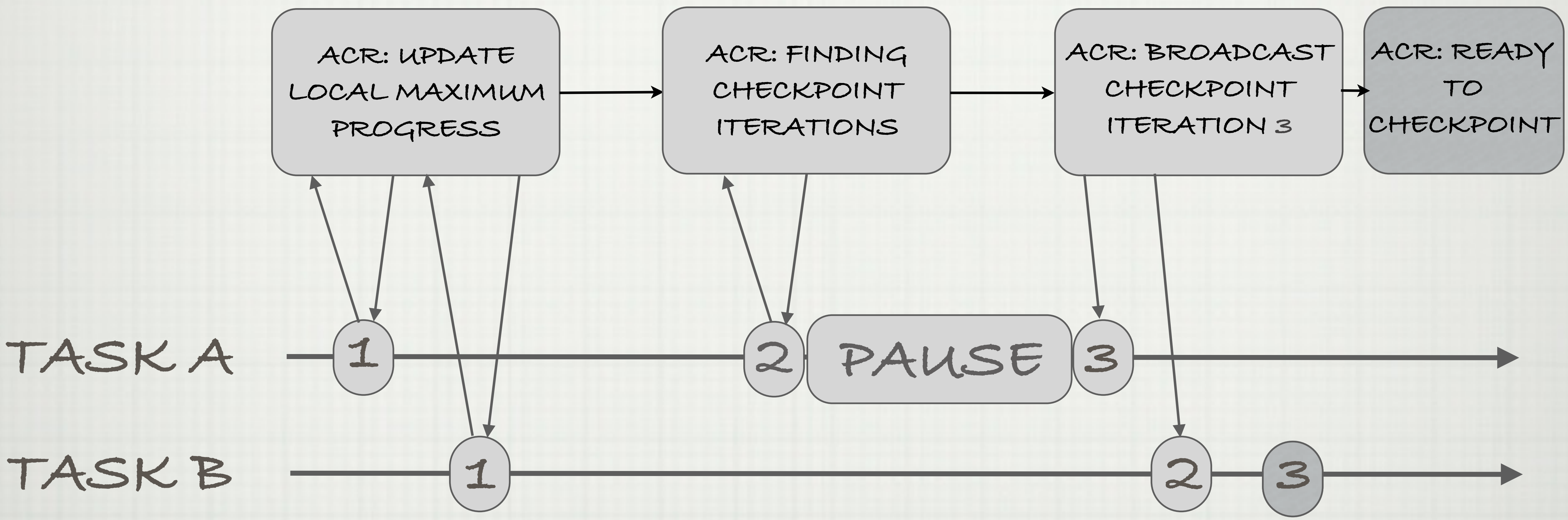
AUTOMATIC CHECKPOINT DECISION



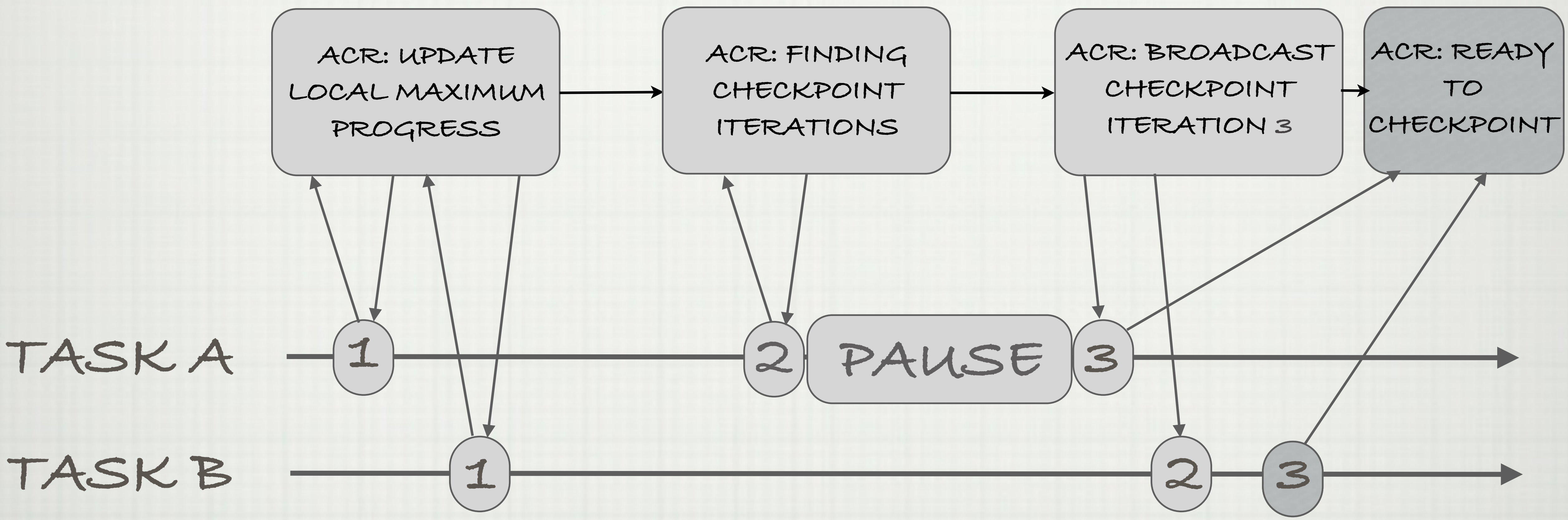
AUTOMATIC CHECKPOINT DECISION



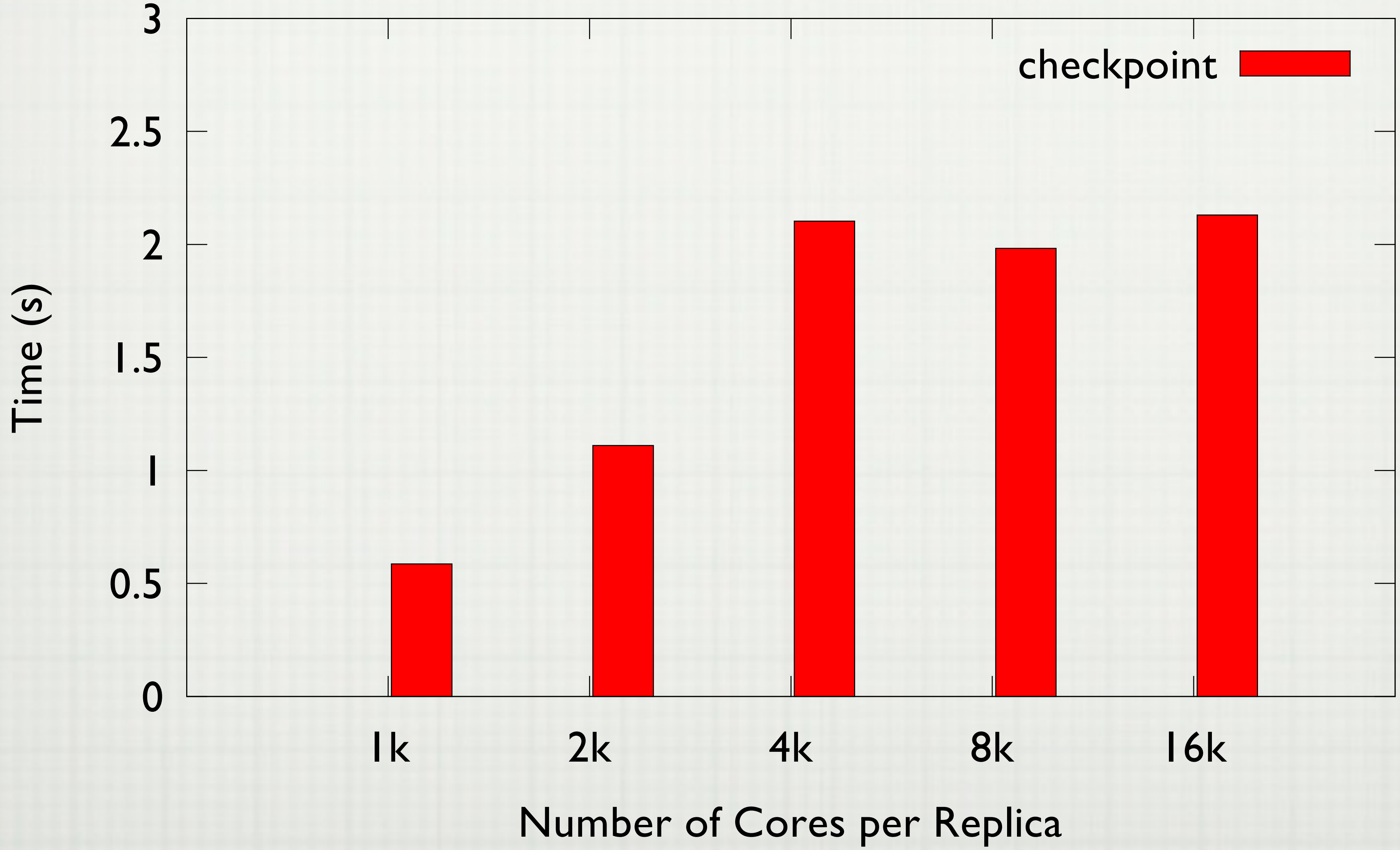
AUTOMATIC CHECKPOINT DECISION



AUTOMATIC CHECKPOINT DECISION

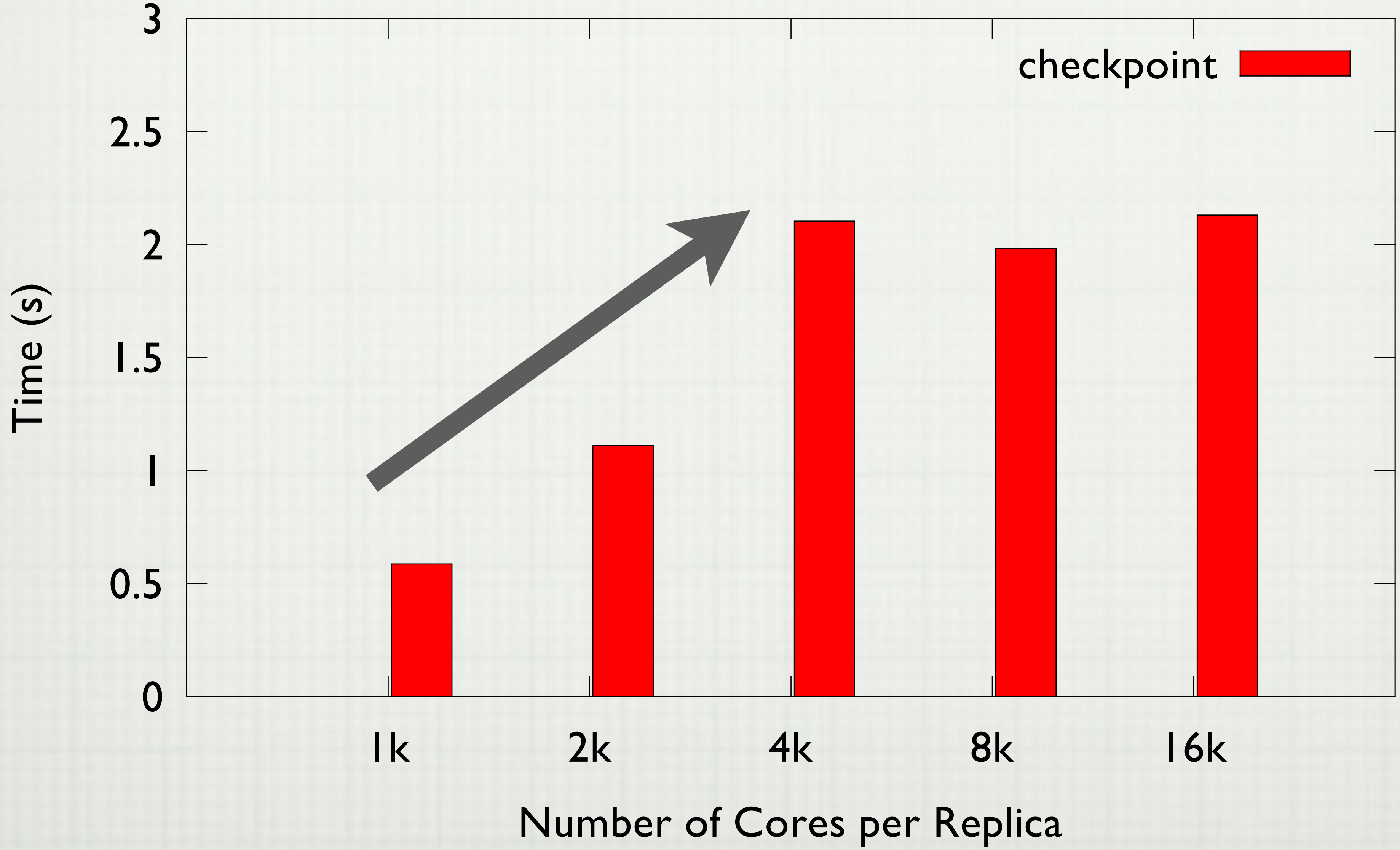


BASE PERFORMANCE



JACOBI3D BGP

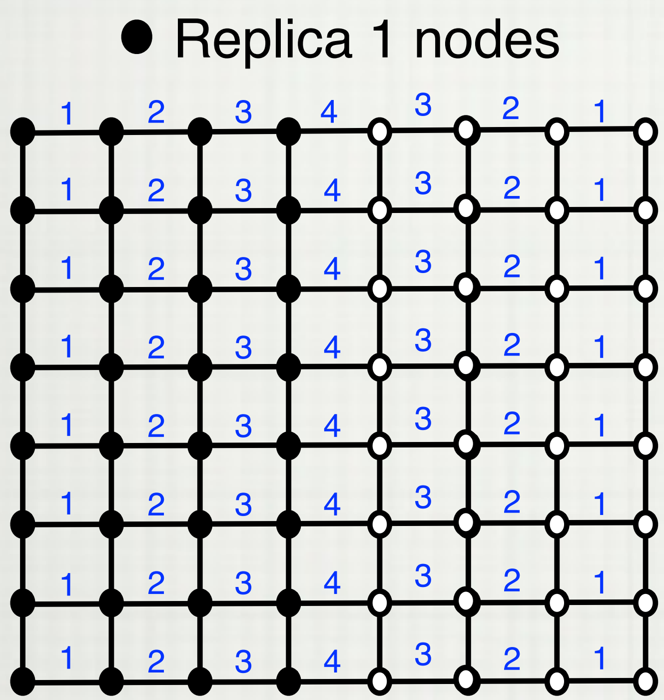
BASE PERFORMANCE



JACOBI3D BGP

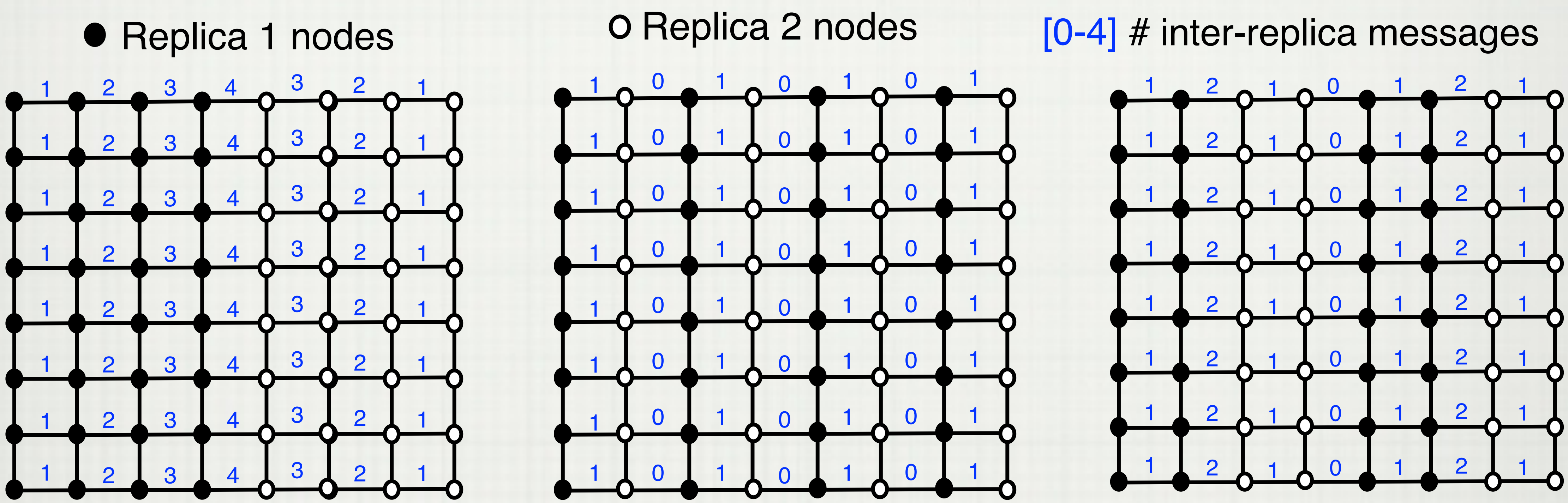
OPTIMIZATION: TOPOLOGY AWARE MAPPING

OPTIMIZATION: TOPOLOGY AWARE MAPPING



(a) Default-mapping

OPTIMIZATION: TOPOLOGY AWARE MAPPING



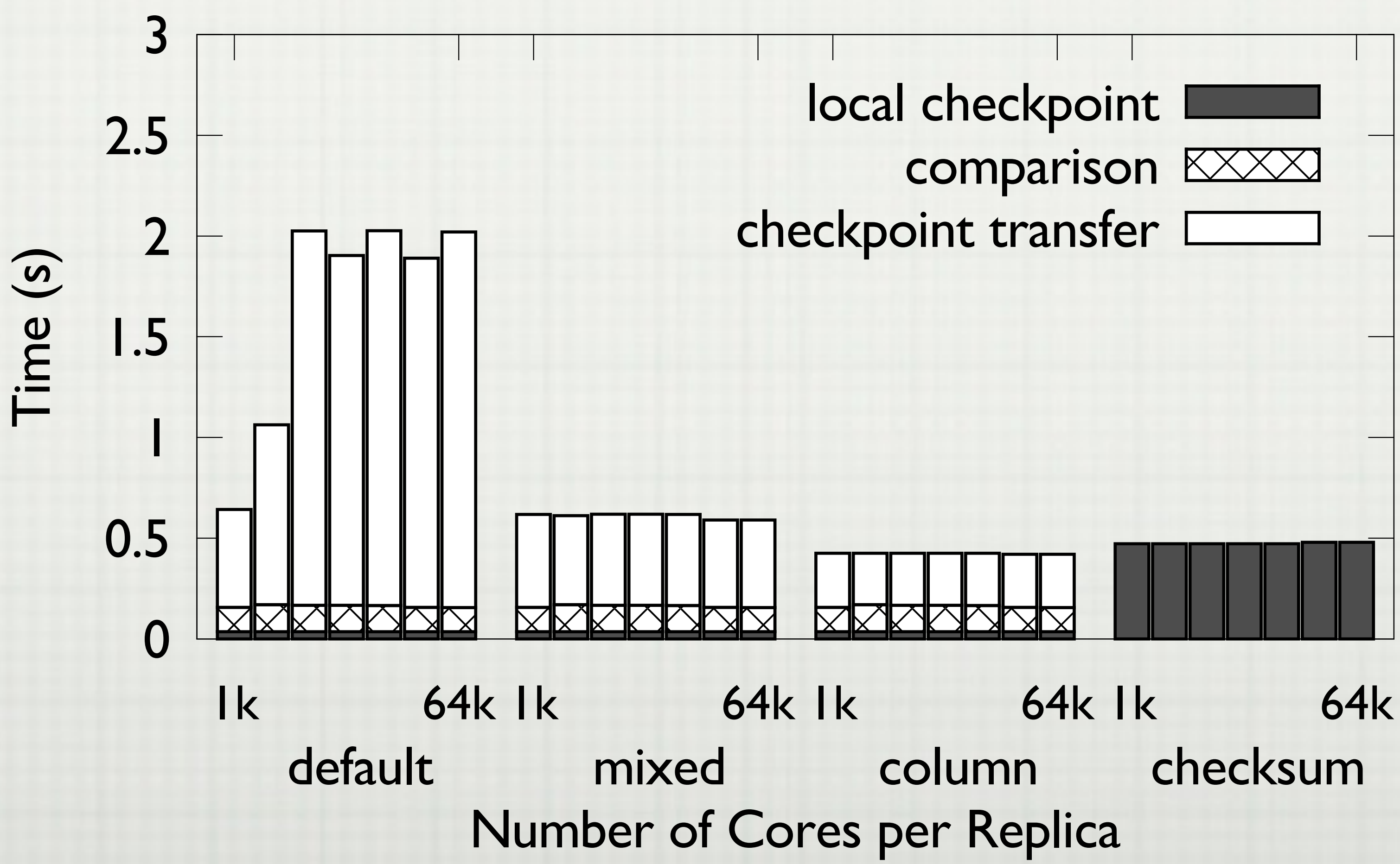
(a) Default-mapping (b) Column-mapping (c) Mixed-mapping

1) Reduce the inter-replica communication distance.

2) Trade-off between inter and intra replica communication.

OPTIMIZATION: CHECKSUM

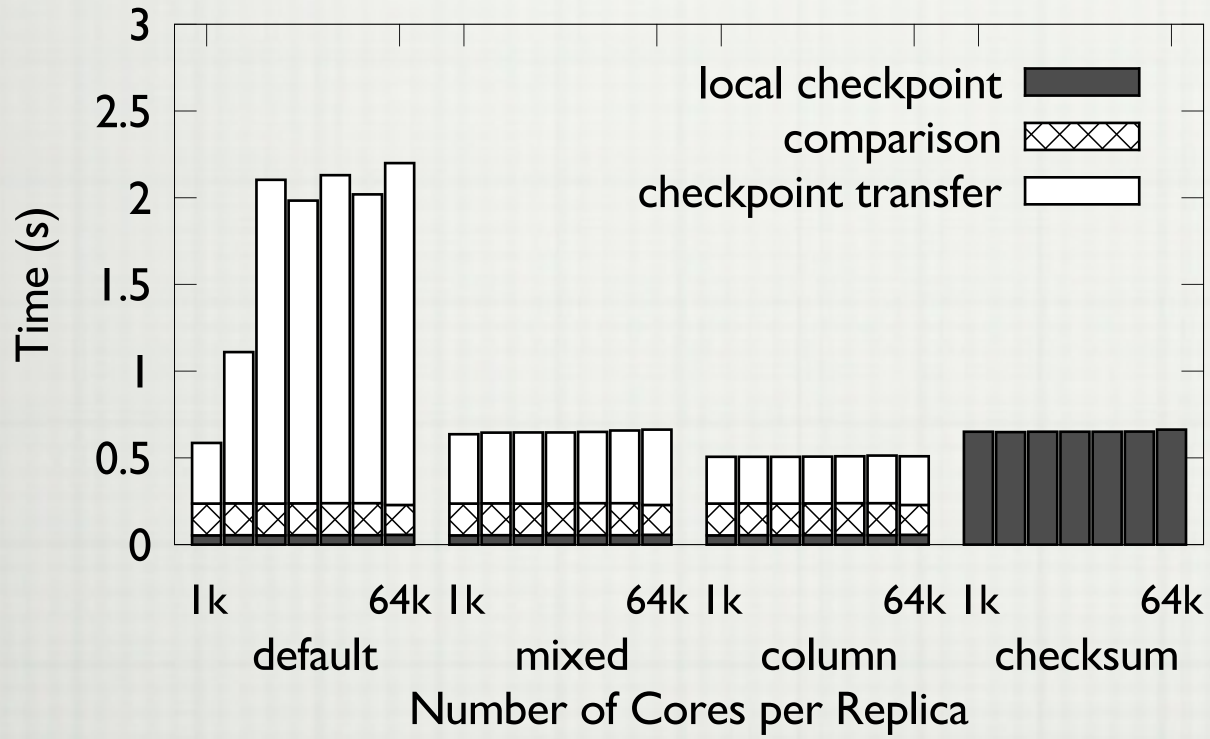
- TRANSFER THE CHECKSUM OF 1 INTEGER INSTEAD OF THE WHOLE CHECKPOINTS
- FLOATING POINT ROUND-OFF ERROR



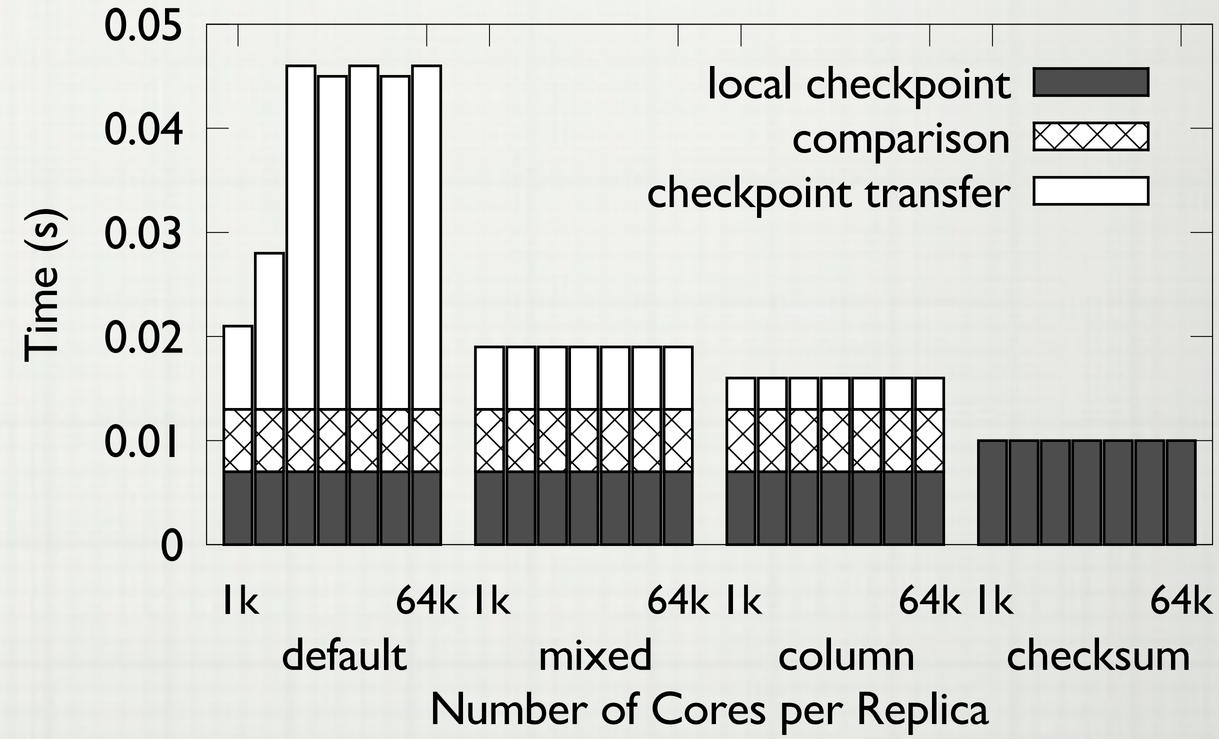
EXPERIMENTAL RESULTS: MINI-APPLICATIONS

Benchmark	Description	Configuration per core	Memory Pressure
Jacobi3D	7-point stencil	64*64*128	High
HPCCG	Unstructured implicit finite element method	40*40*40	High
LULESH	Unstructured explicit mesh	32*32*64	High
LeanMD	Short-range non-bonded force calculation in NAMD	4000 atoms	Low
miniMD	Mimic the performance in LAMMPS	1000 atoms	Low

EXPERIMENTAL RESULTS: CHECKPOINT

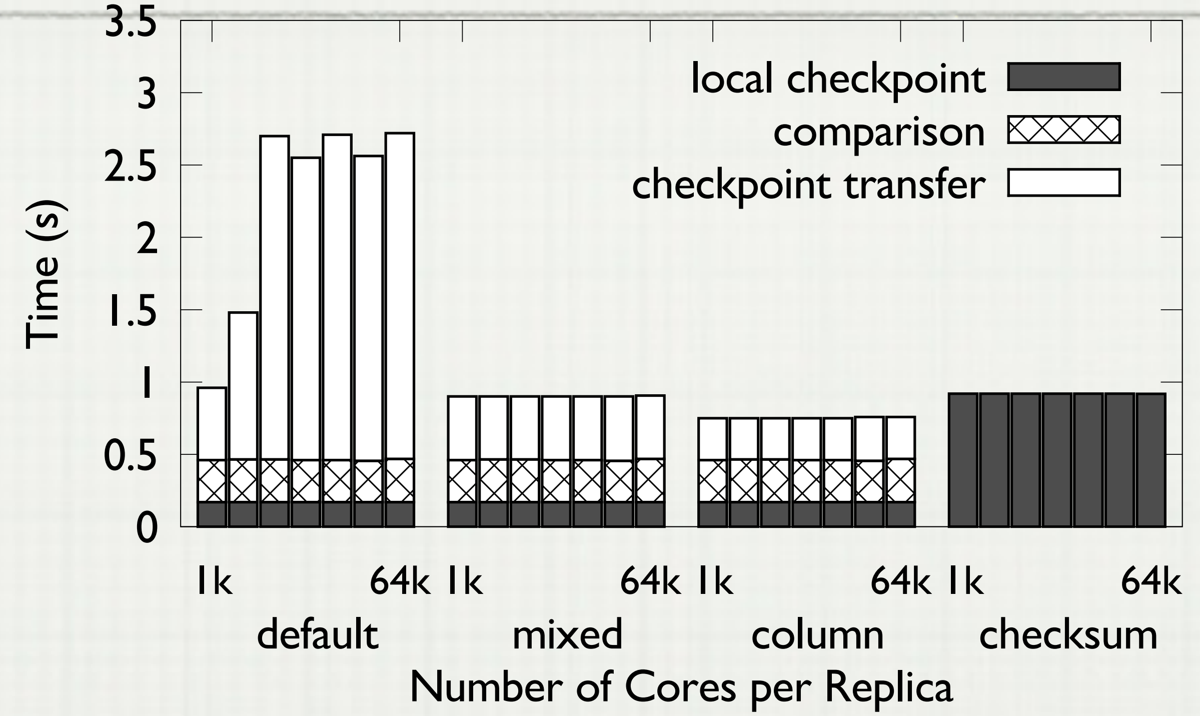


JACOBI3D AMPI

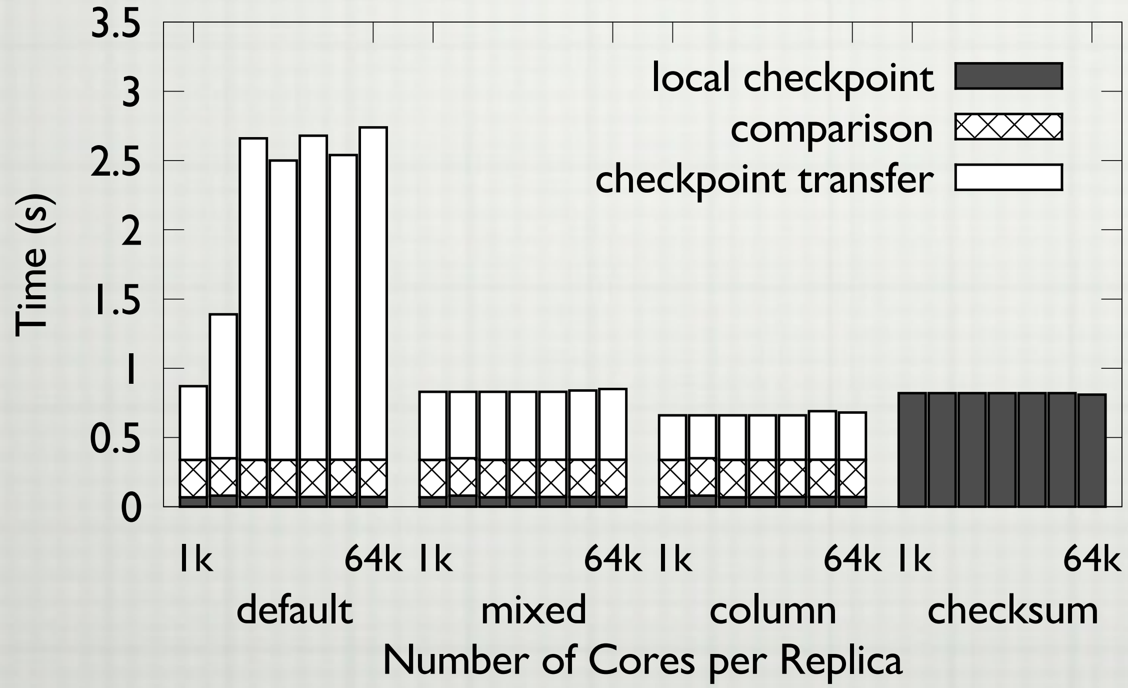


LEANMD

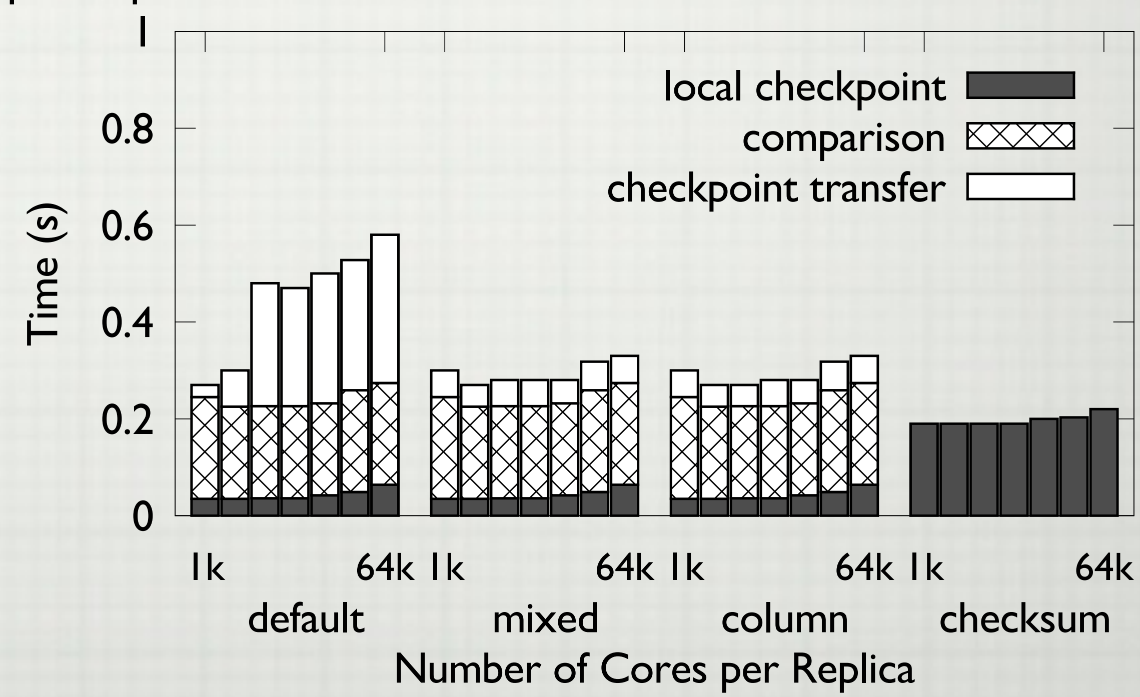
EXPERIMENTAL RESULTS: CHECKPOINT



LULESH

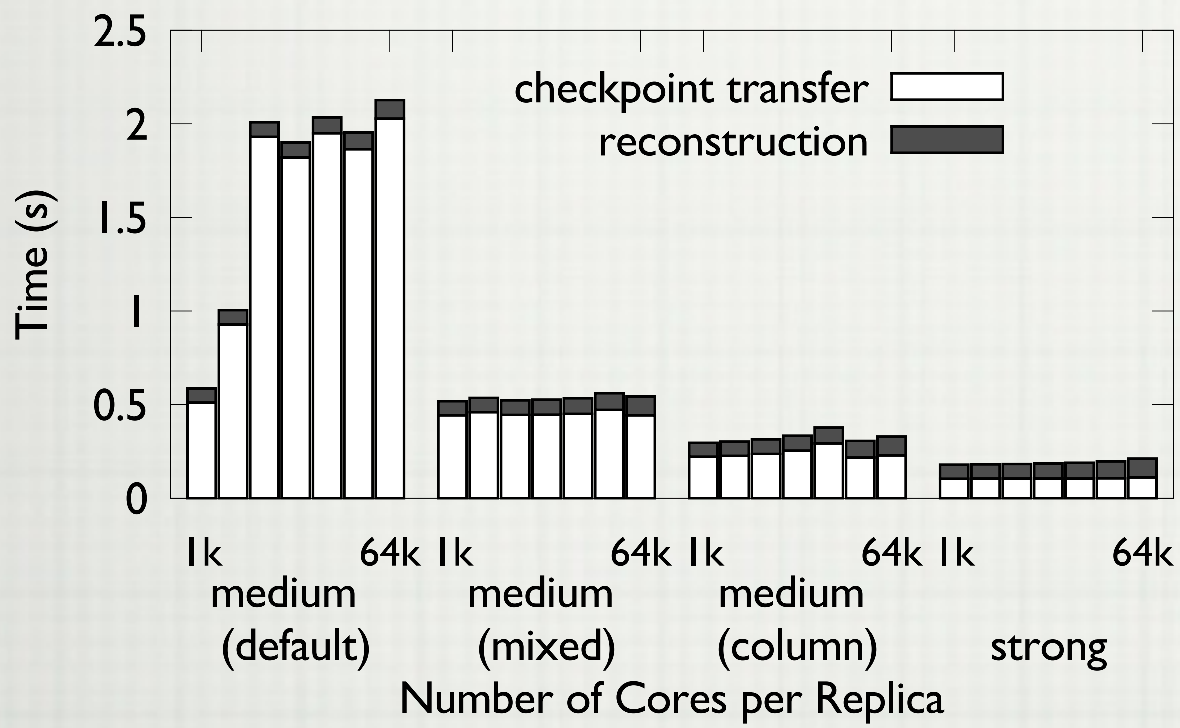


HPCCG

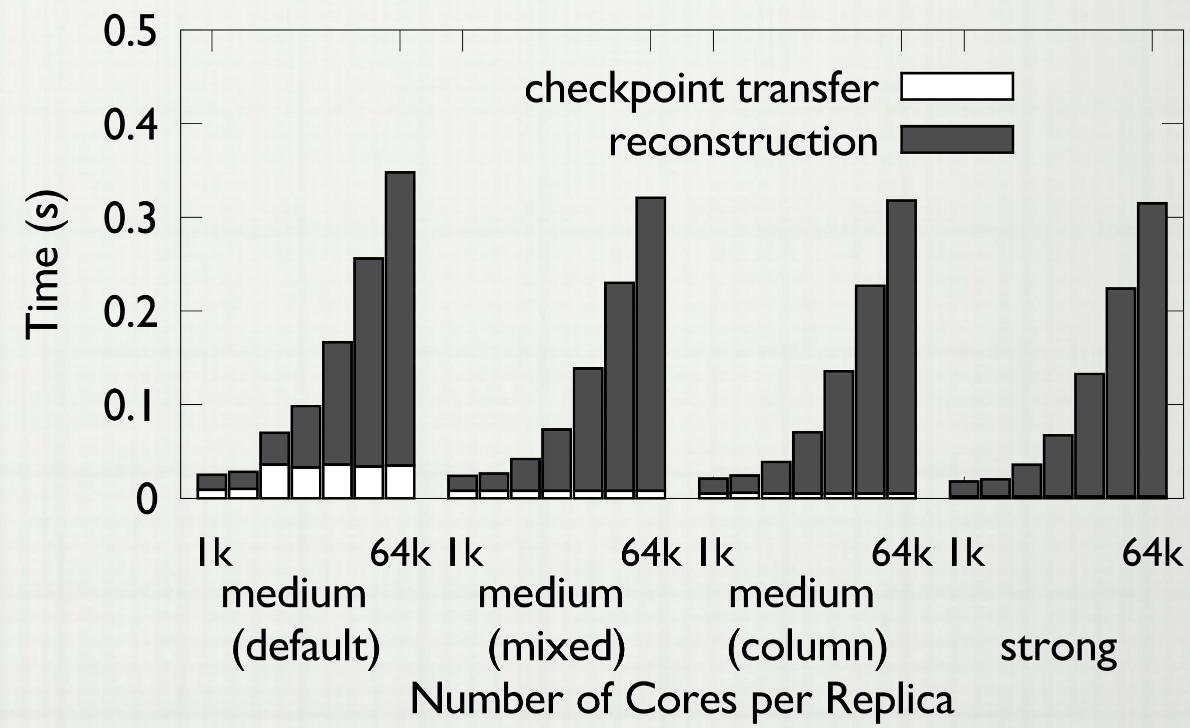


MINIMD

EXPERIMENTAL RESULTS: RESTART

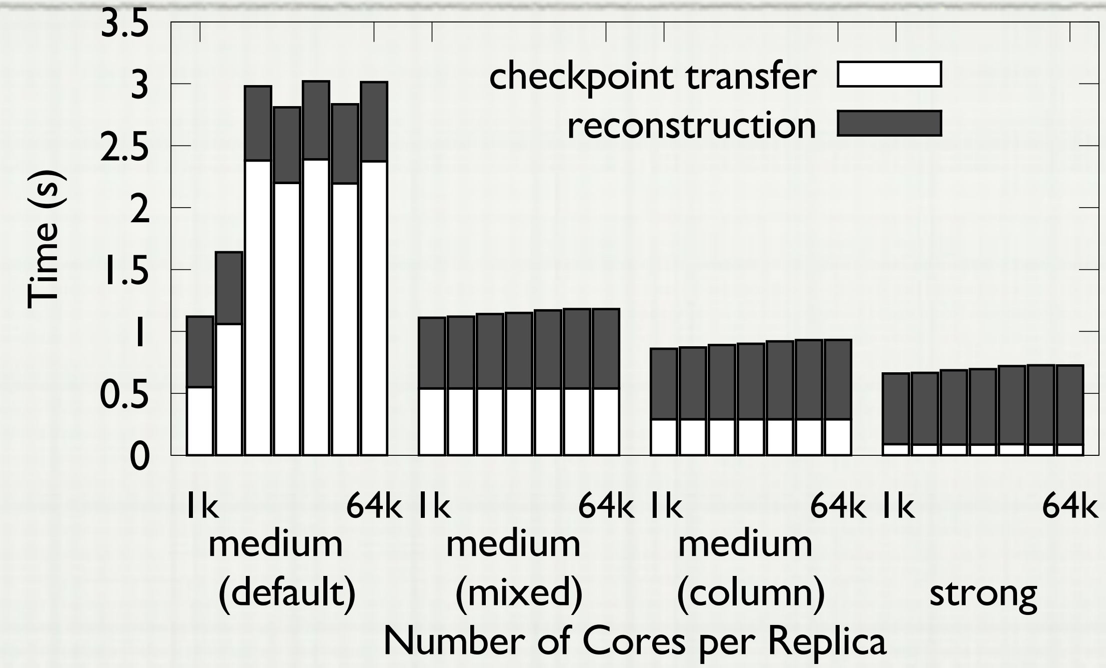


JACOBI3D AMPI

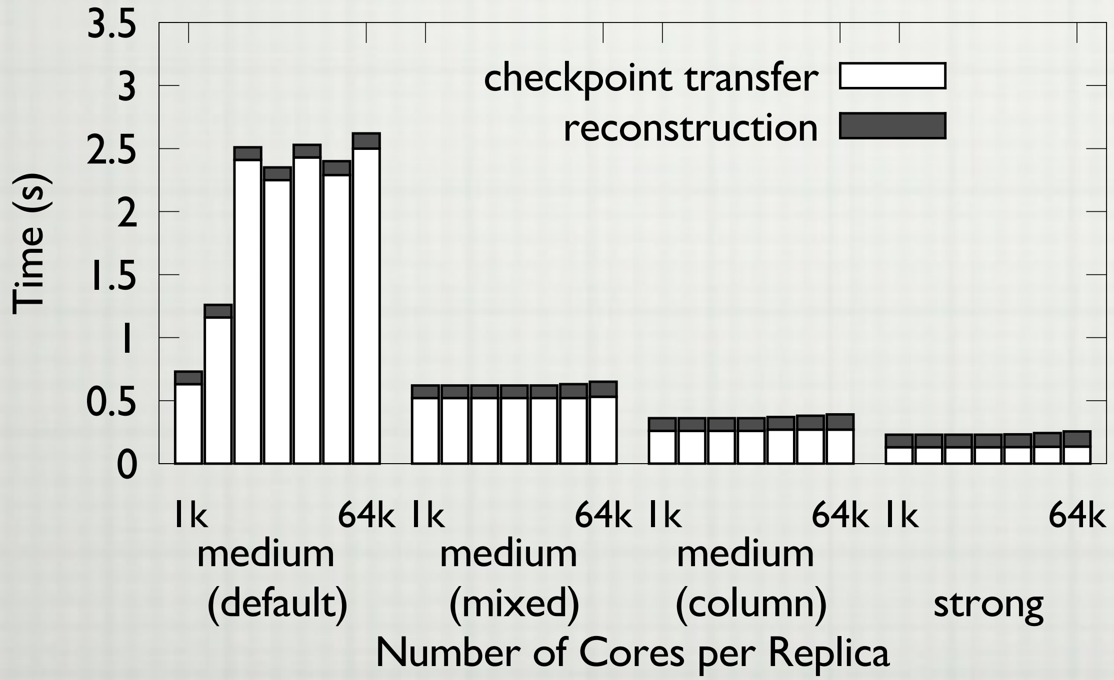


LEANMD

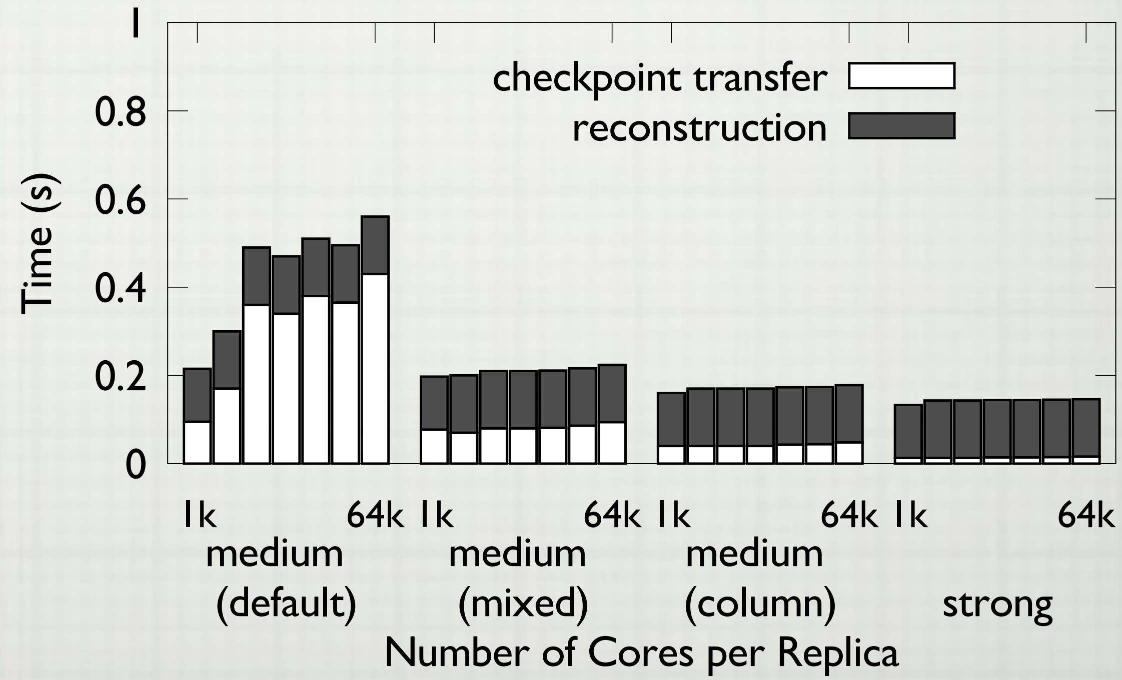
EXPERIMENTAL RESULTS: RESTART



LULESH

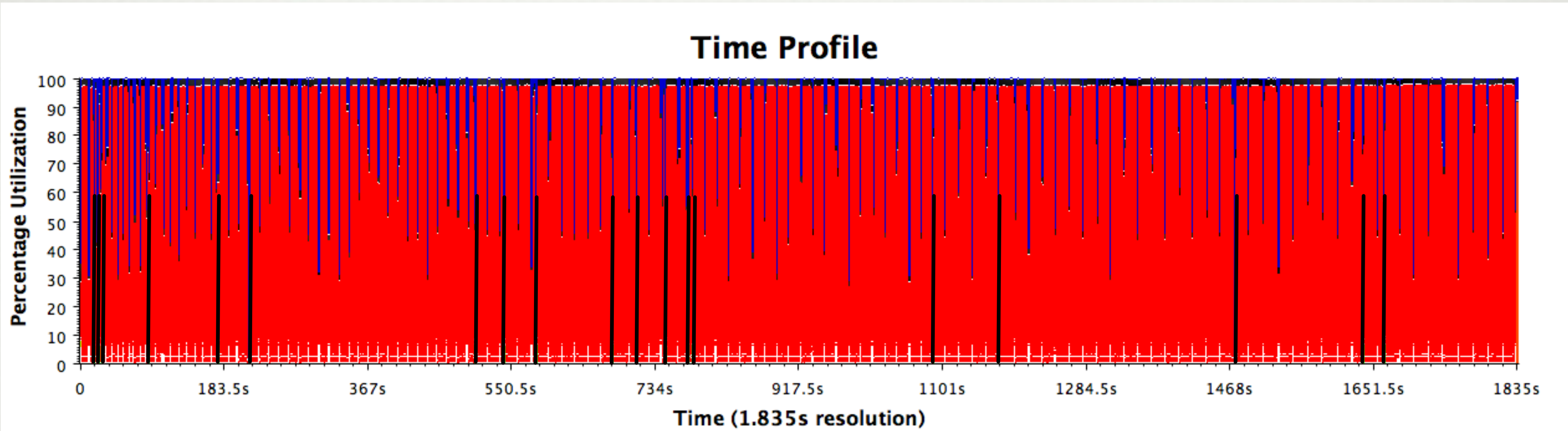


HPCCG



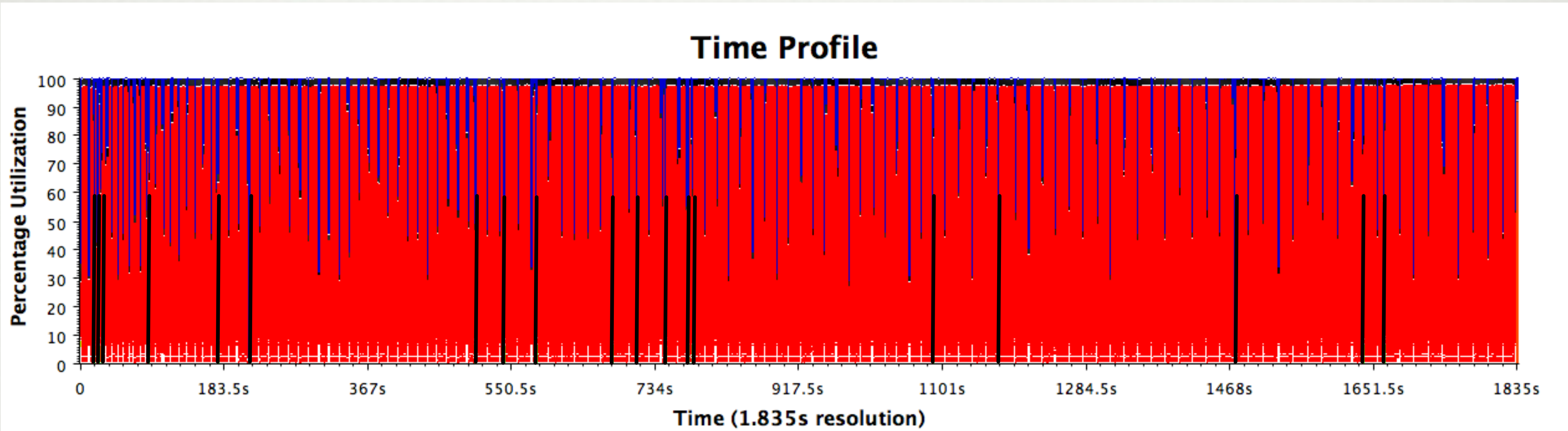
MINIMD

ADAPTING TO FAILURES



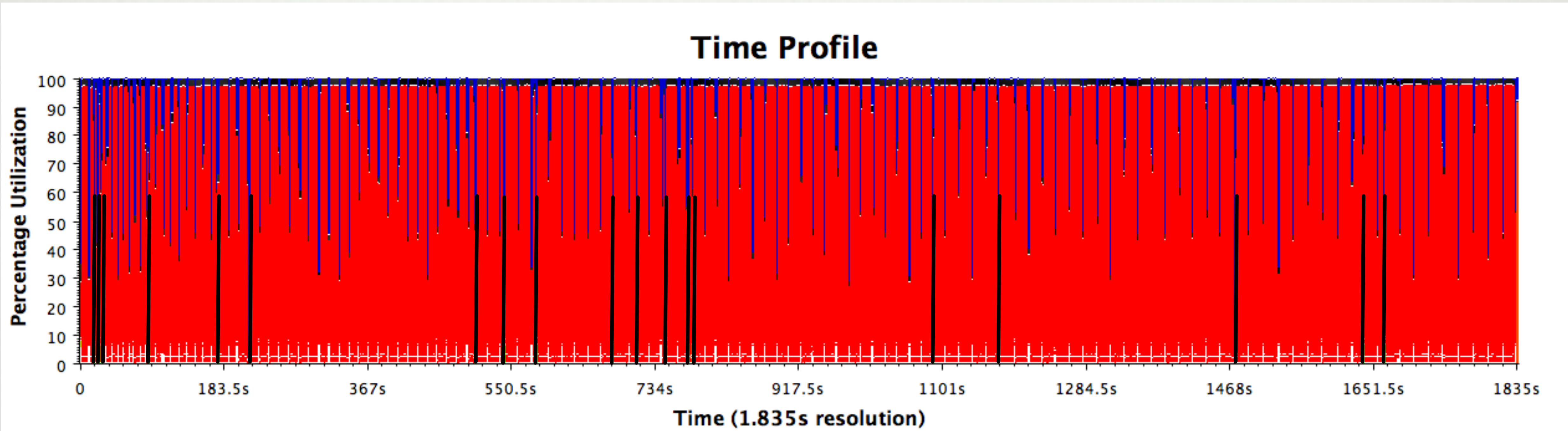
- Failures are injected according to Weibull process: shape parameter 0.6
- 19 failures are injected in a 30mins Jacobi3d run. 14 failures are injected in the first half while 5 failures are injected in the second half.
- Real failures injected: node becomes irresponsive
- Automatic restart: with the support of spare nodes and thus no need to submit the job again

ADAPTING TO FAILURES



- Runtime system observes the change in failure rate based on the failure history.
- Adjusts the optimum checkpoint period based on Daly's model.
- Checkpoint period changes from 6s in the beginning to 17s in the end.

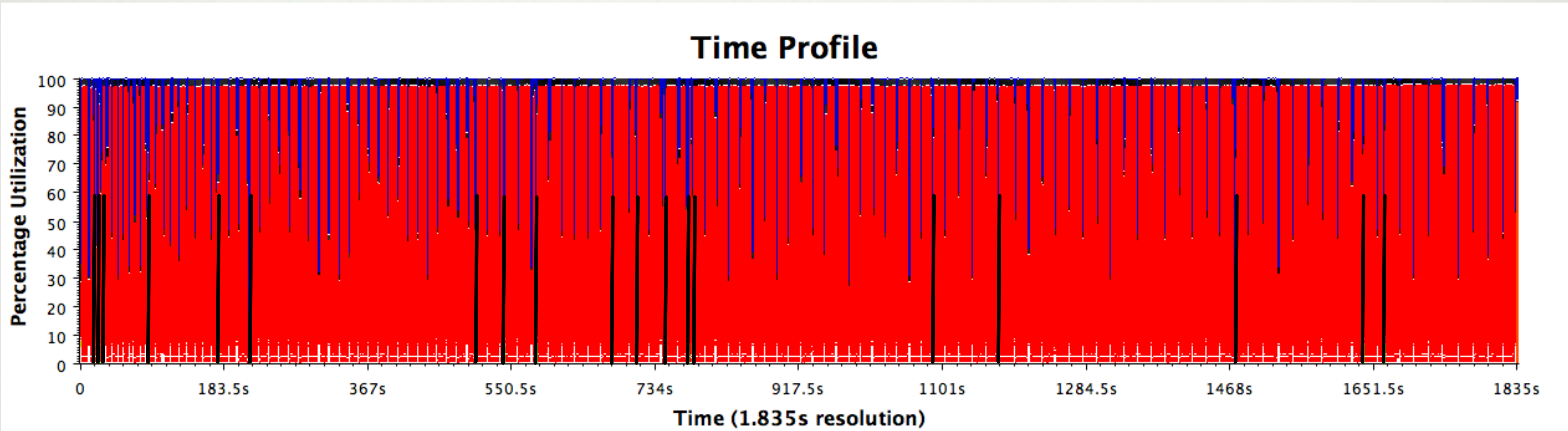
ADAPTING TO FAILURES



6s

- Runtime system observes the change in failure rate based on the failure history.
- Adjusts the optimum checkpoint period based on Daly's model.
- Checkpoint period changes from 6s in the beginning to 17s in the end.

ADAPTING TO FAILURES



6s

- Runtime system observes the change in failure rate based on the failure history.
- Adjusts the optimum checkpoint period based on Daly's model.
- Checkpoint period changes from 6s in the beginning to 17s in the end.

17s

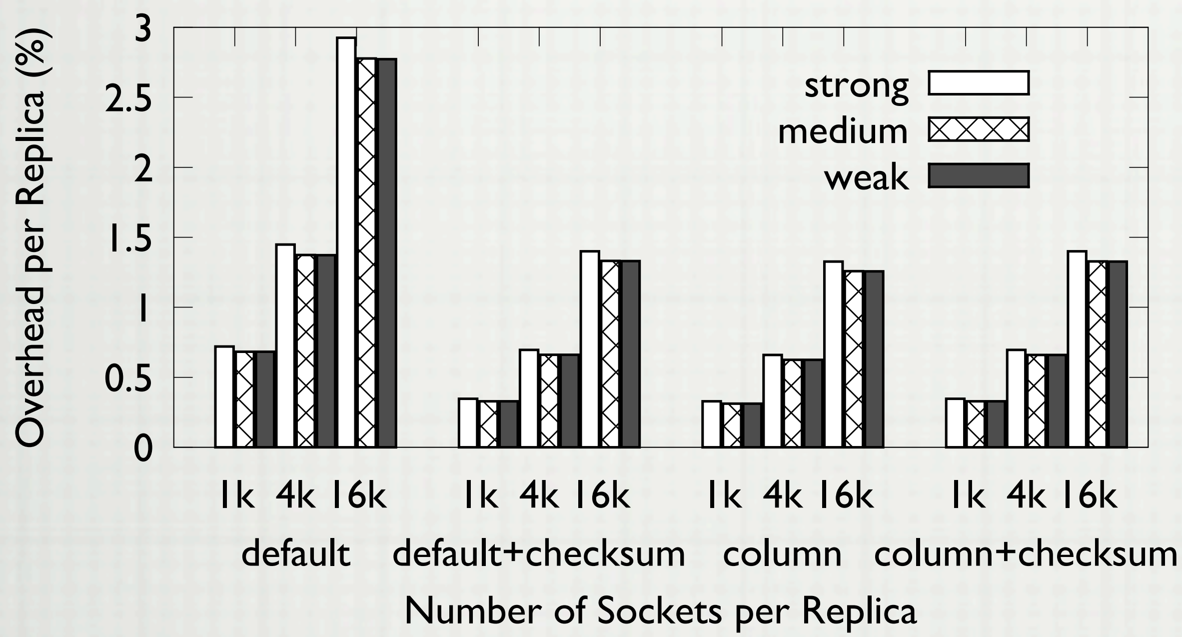
CONCLUSION

- Protection for both hard errors and silent data corruption is important
- Replication enables synergistic handling of errors
- Good scalability
- Automatic checkpoint decision
  - Restart from hard errors
  - Adapting to different failure distributions

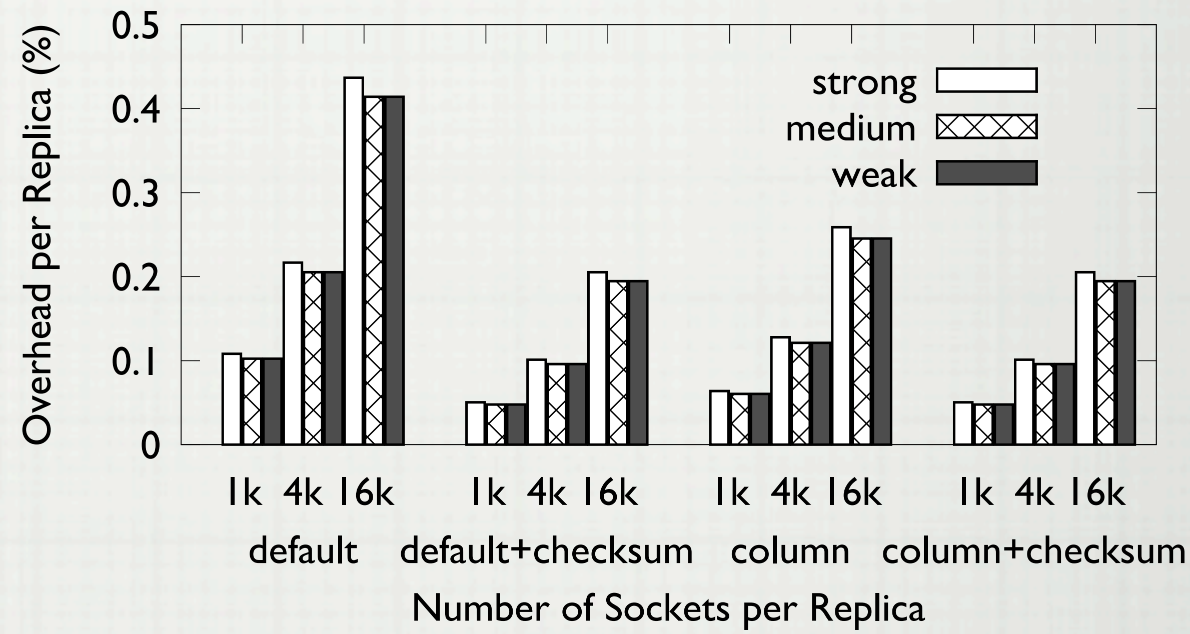
THANKS!

- QUESTIONS?
- CONTACT AT xiangni2@illinois.edu
- MORE INFO AT http://charm.cs.illinois.edu/research/ft

OVERHEAD



JACOBI3D



LEANMD

MODELING OF UTILIZATION AND VULNERABILITY