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

XIANG NI, ESTEBAN MENESES, NIKHIL JAIN, SANJAY KALE

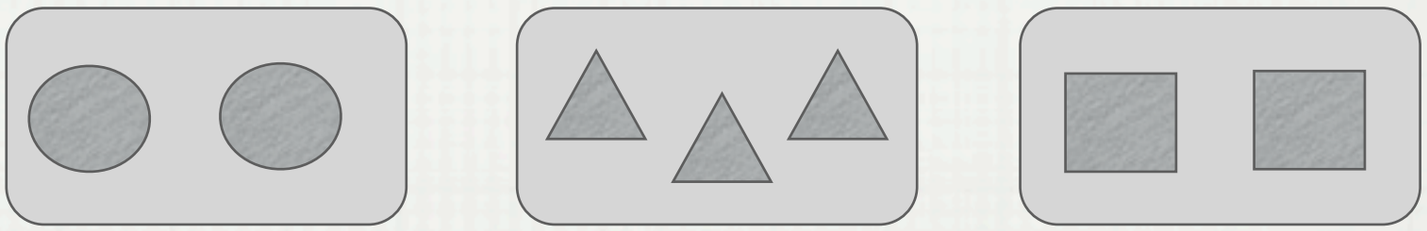
PARALLEL PROGRAMMING LAB, UIUC

CONTENTS

- ☐ MOTIVATION
- ☐ ACR FRAMEWORK
- ☐ OPTIMIZATION
- ☐ EXPERIMENTAL RESULTS
- ☐ CONCLUSION

BACKGROUND

TASK

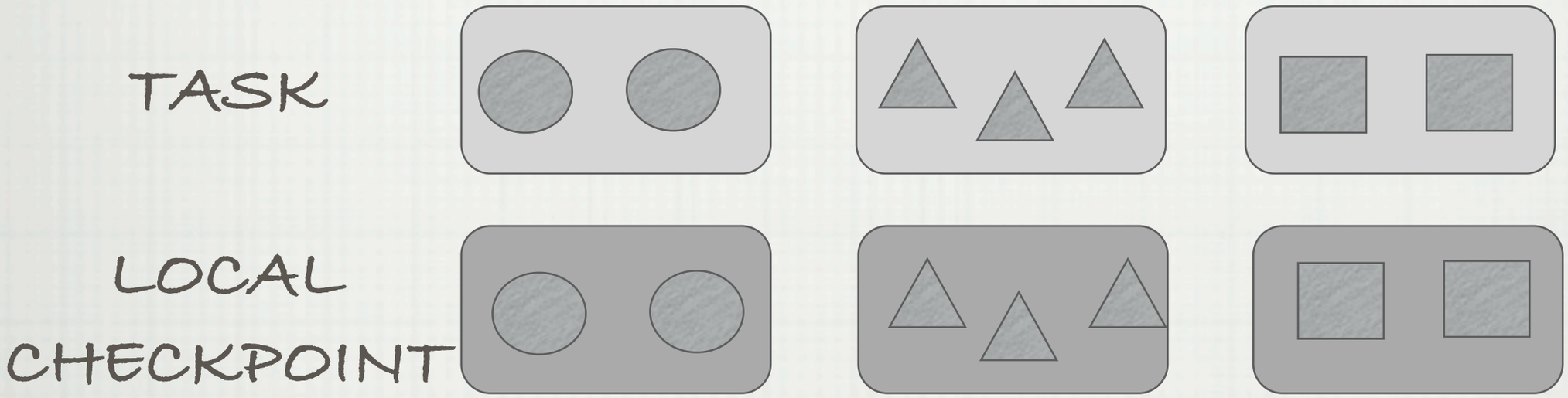


NODE A

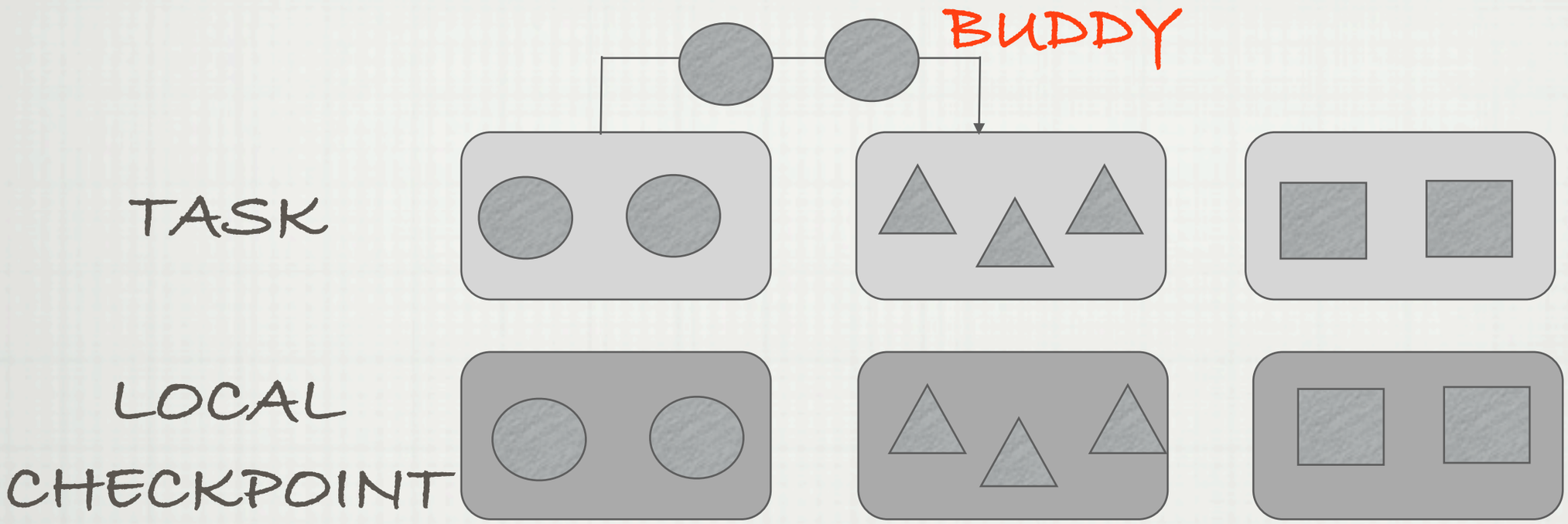
NODE B

NODE C

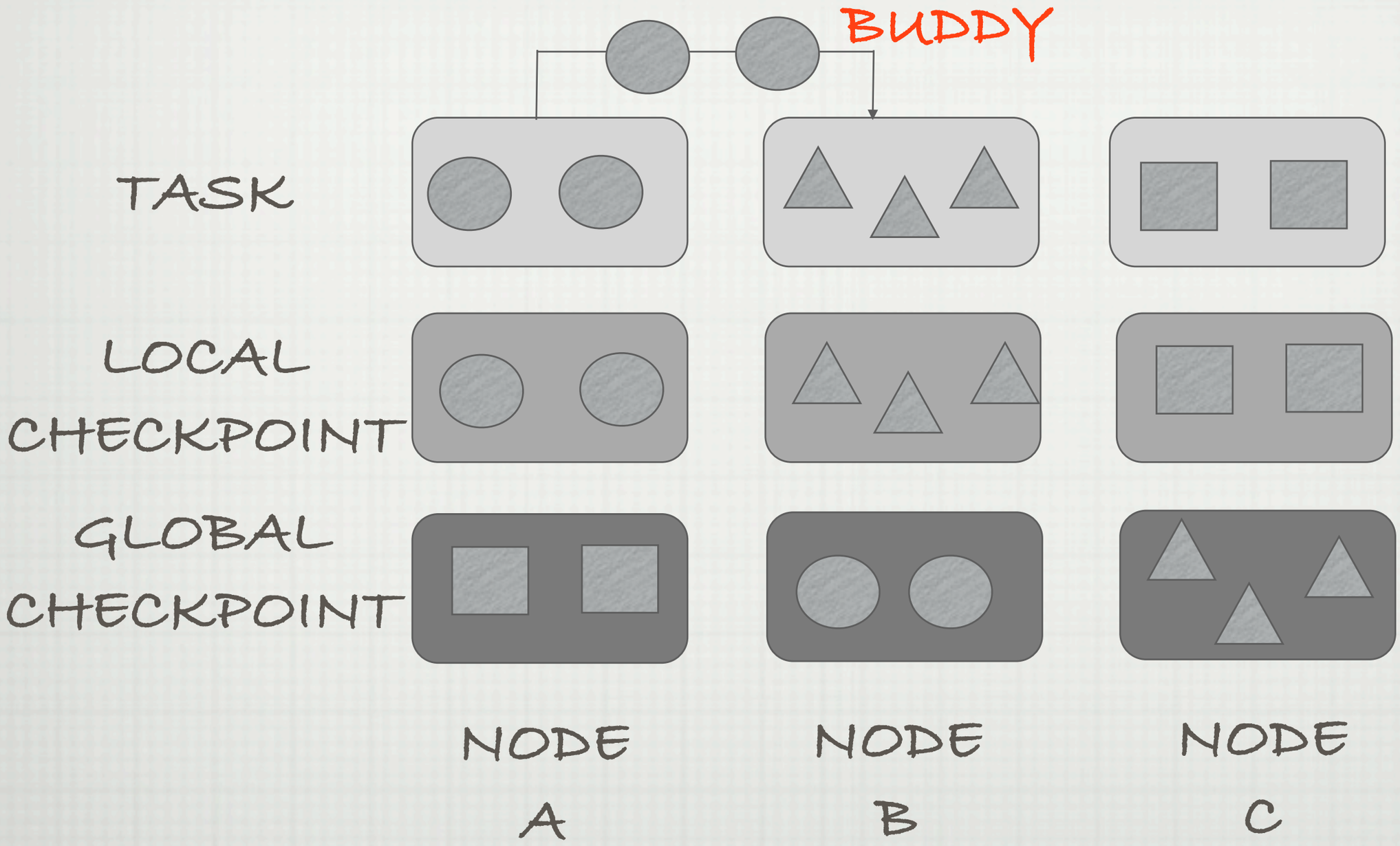
BACKGROUND



BACKGROUND



BACKGROUND


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

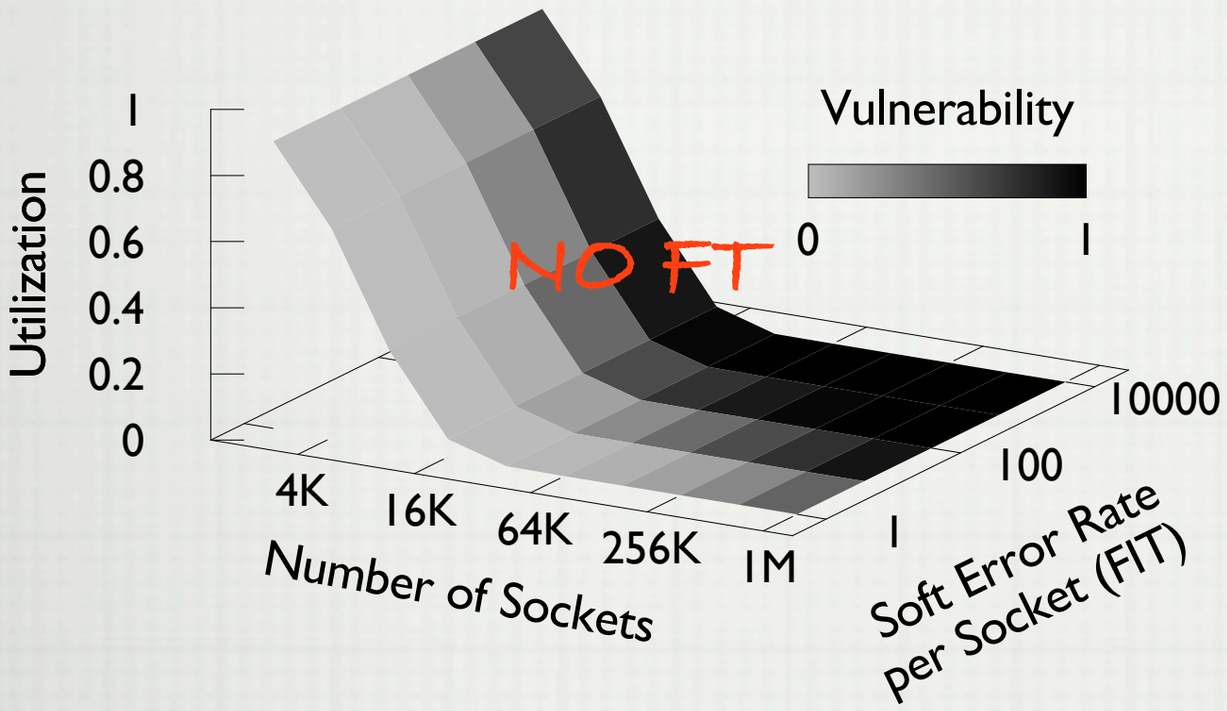
MOTIVATION

- New challenge: soft error.
- 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?

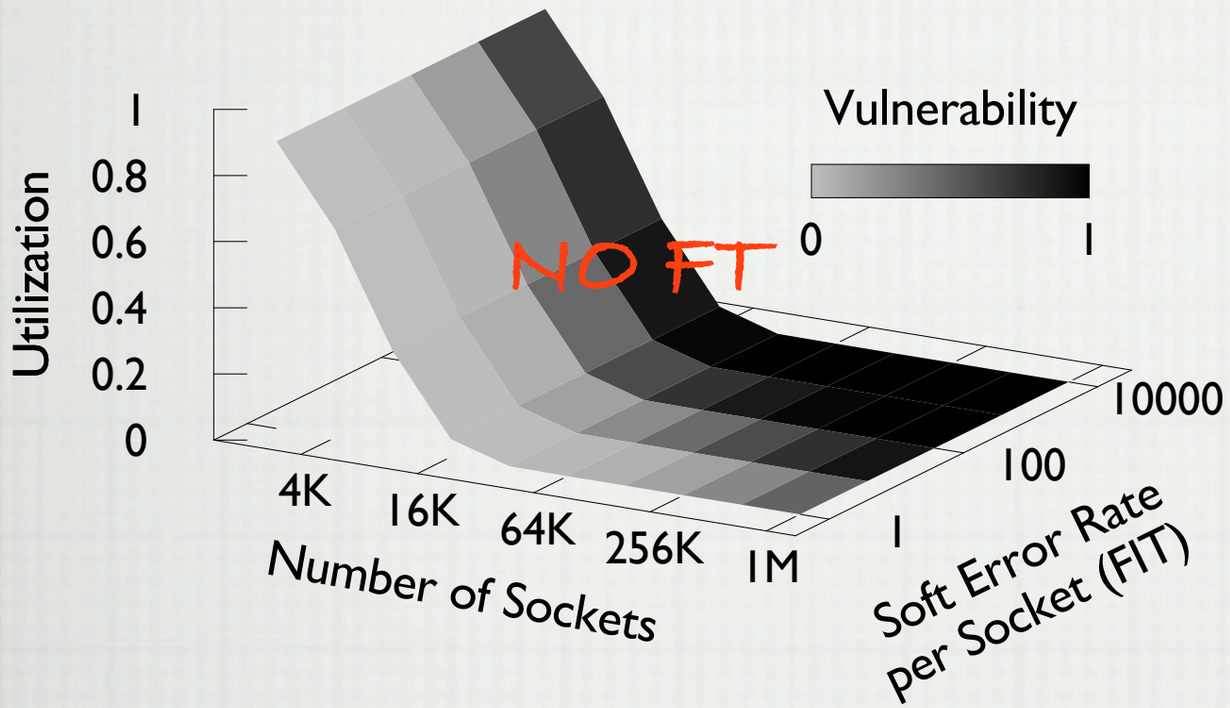
MOTIVATION

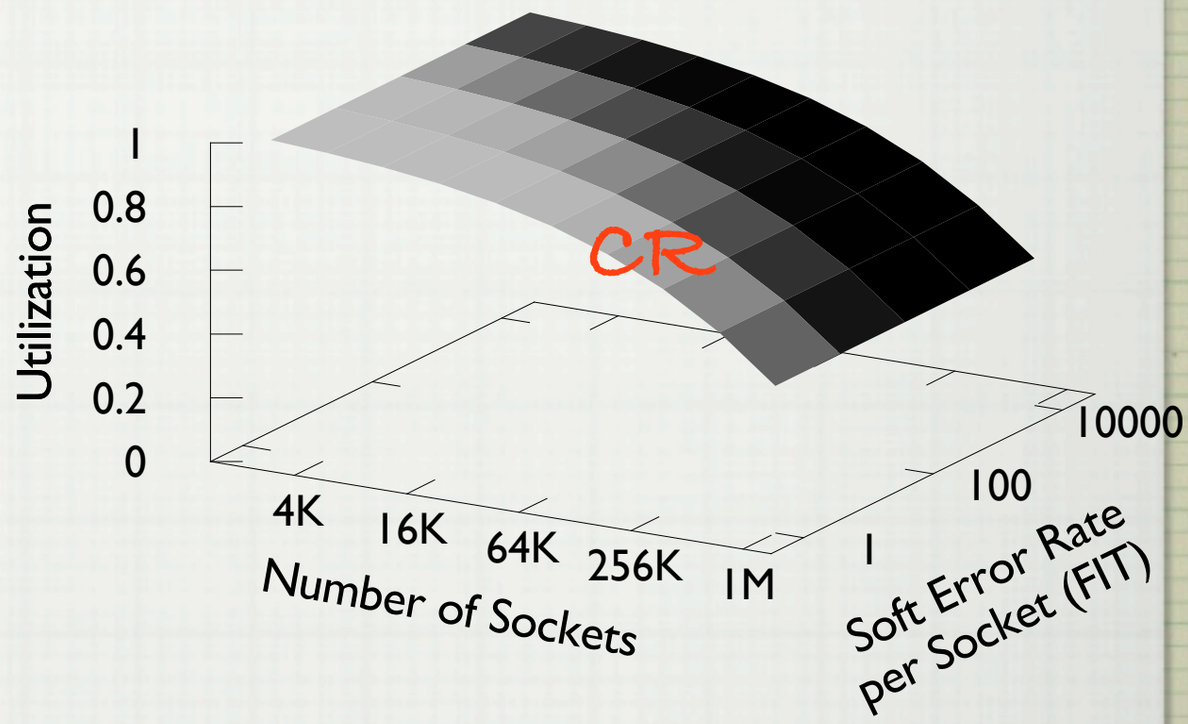
MOTIVATION

MOTIVATION

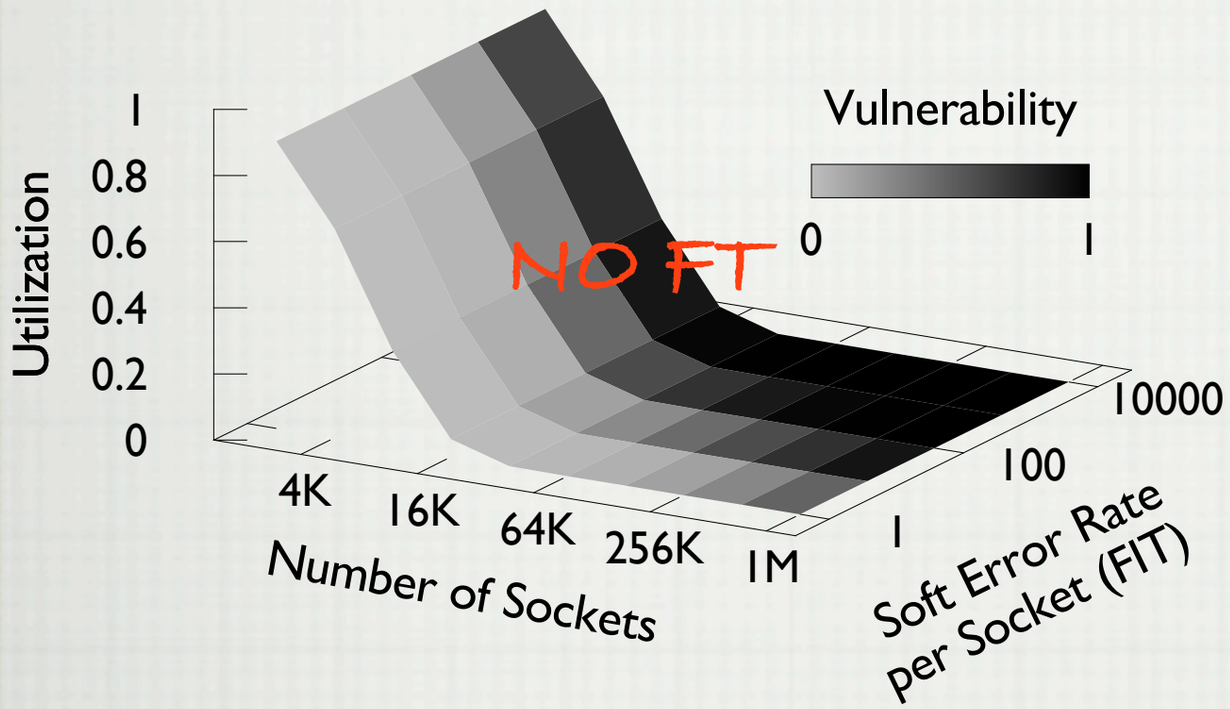

Utilization
1
0.8
0.6
0.4
0.2
0
Vulnerability
0
1
NO FT
4K
16K
64K
256K
1M
Number of Sockets
1
100
10000
Soft Error Rate
per Socket (FIT)

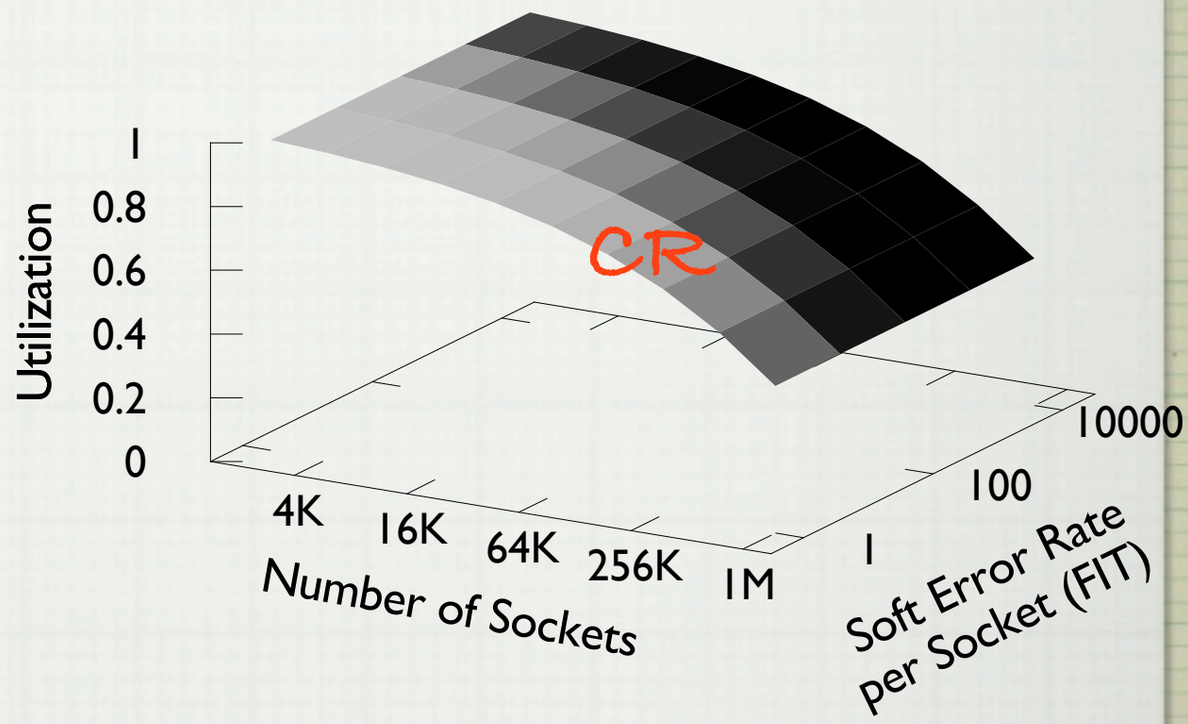
MOTIVATION

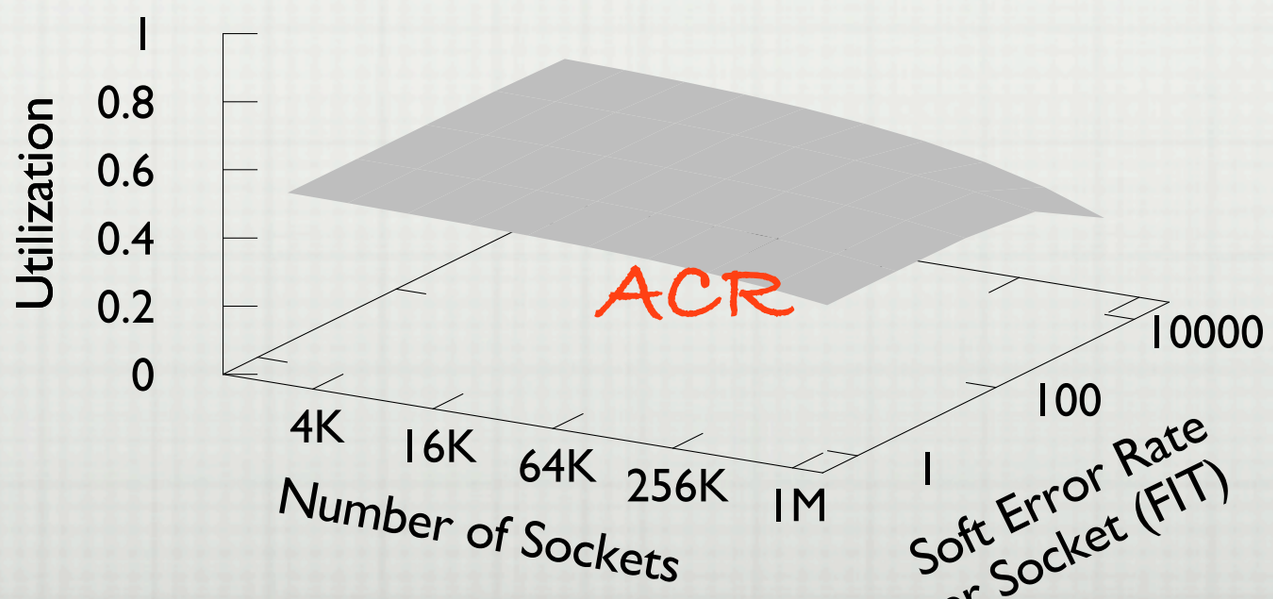

Utilization
1
0.8
0.6
0.4
0.2
0
Vulnerability
0
1
NO FT
4K
16K
64K
256K
1M
Number of Sockets
1
100
10000
Soft Error Rate
per Socket (FIT)


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

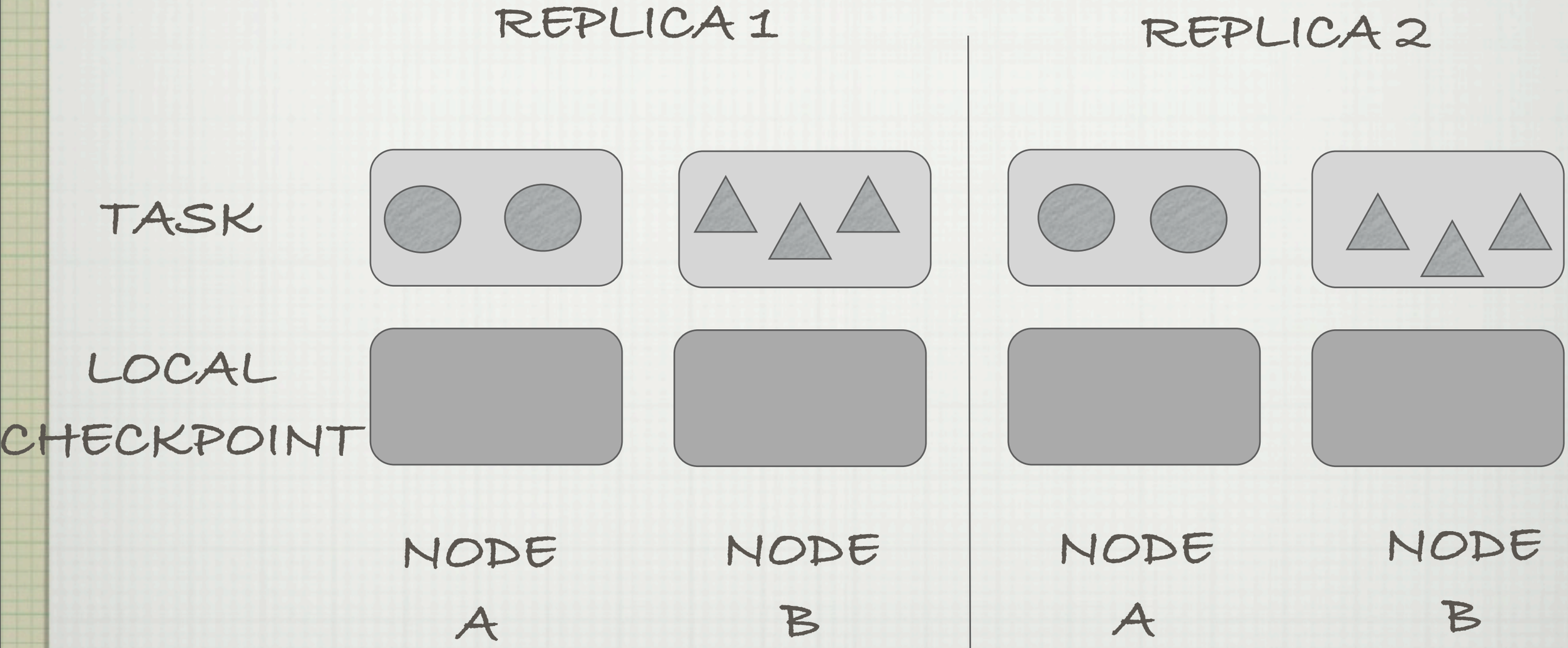
MOTIVATION


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

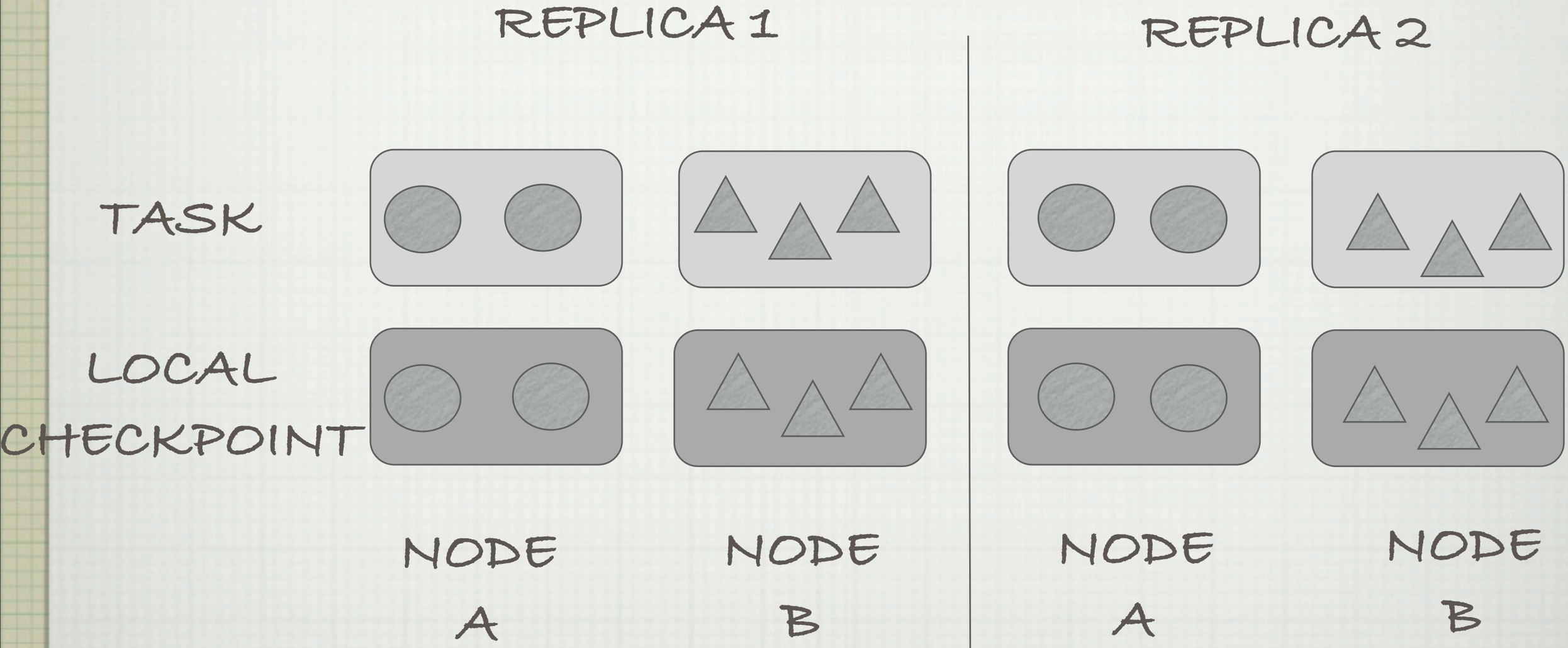

Utilization
1
0.8
0.6
0.4
0.2
0
CR
4K
16K
64K
256K
1M
Number of Sockets
1
100
10000
Soft Error Rate
per Socket (FIT)


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

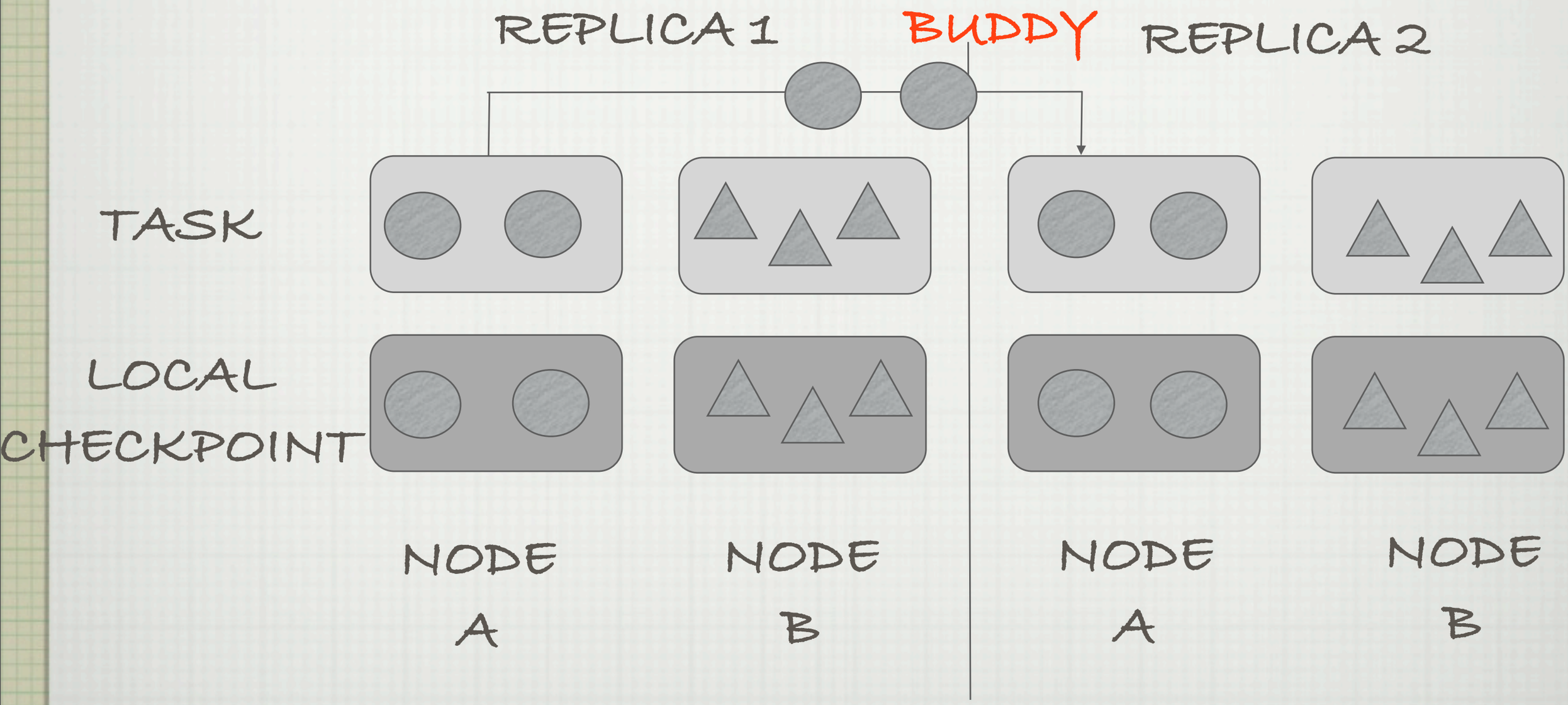
REPLICATION ENHANCED CHECKPOINTING



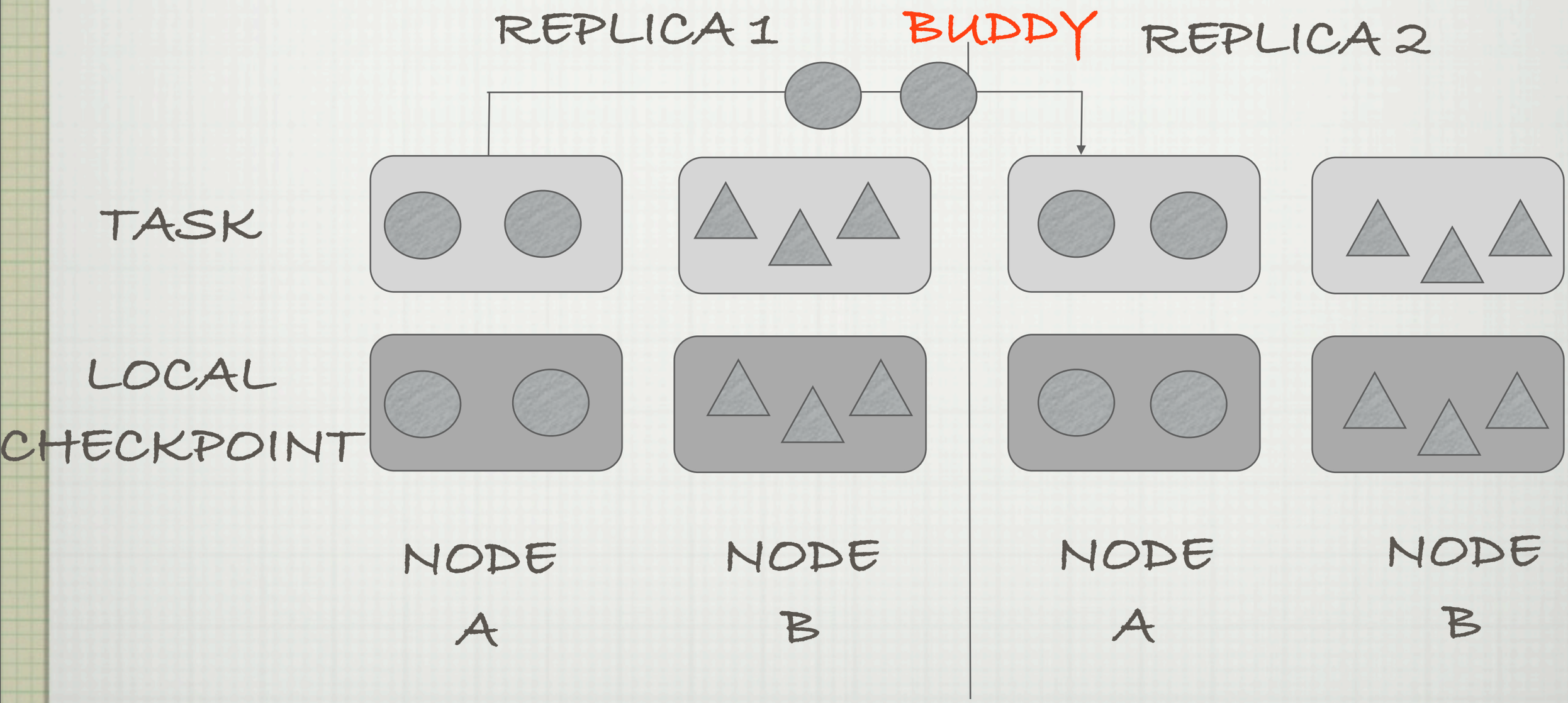
REPLICATION ENHANCED CHECKPOINTING



REPLICATION ENHANCED CHECKPOINTING

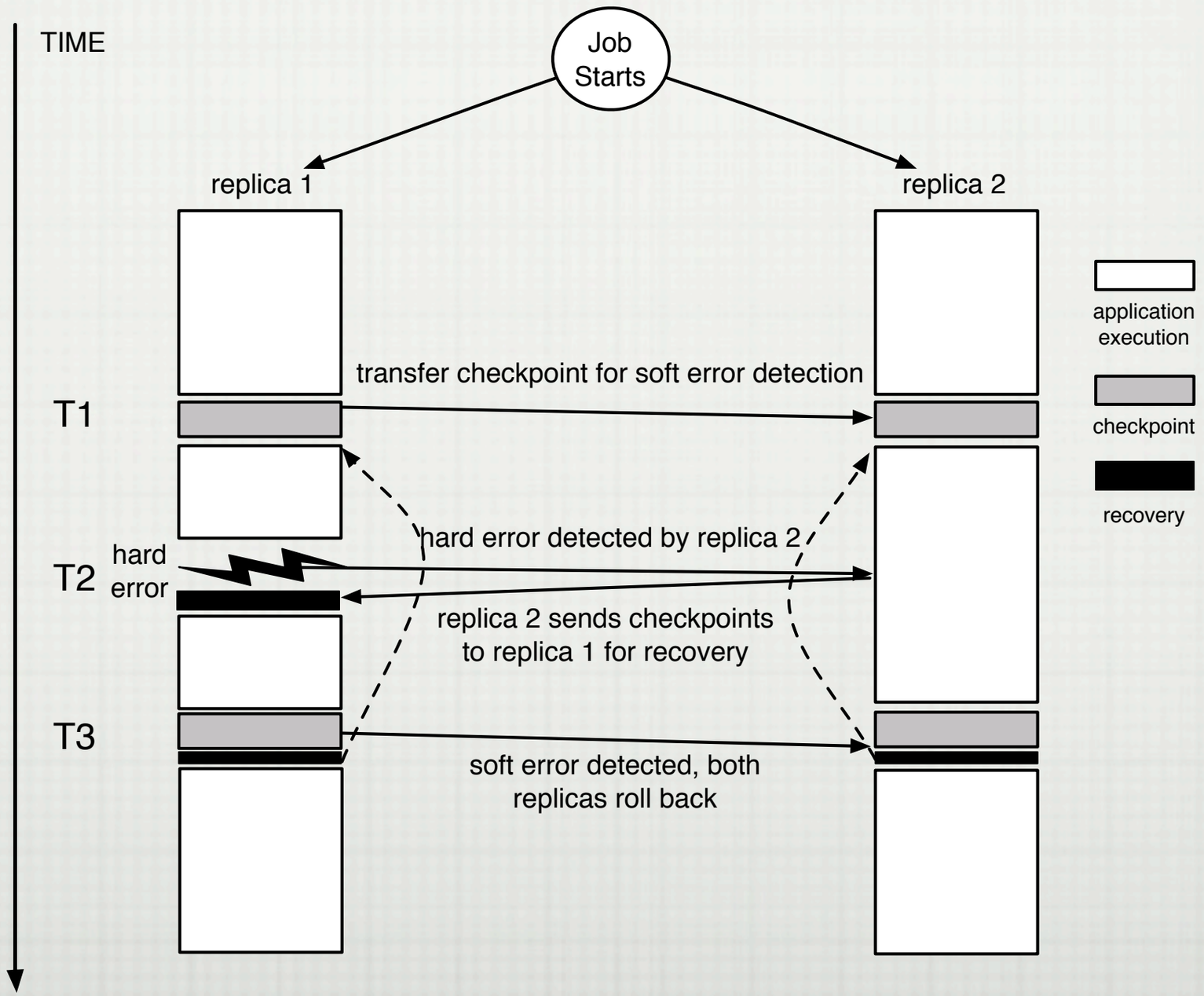


REPLICATION ENHANCED CHECKPOINTING



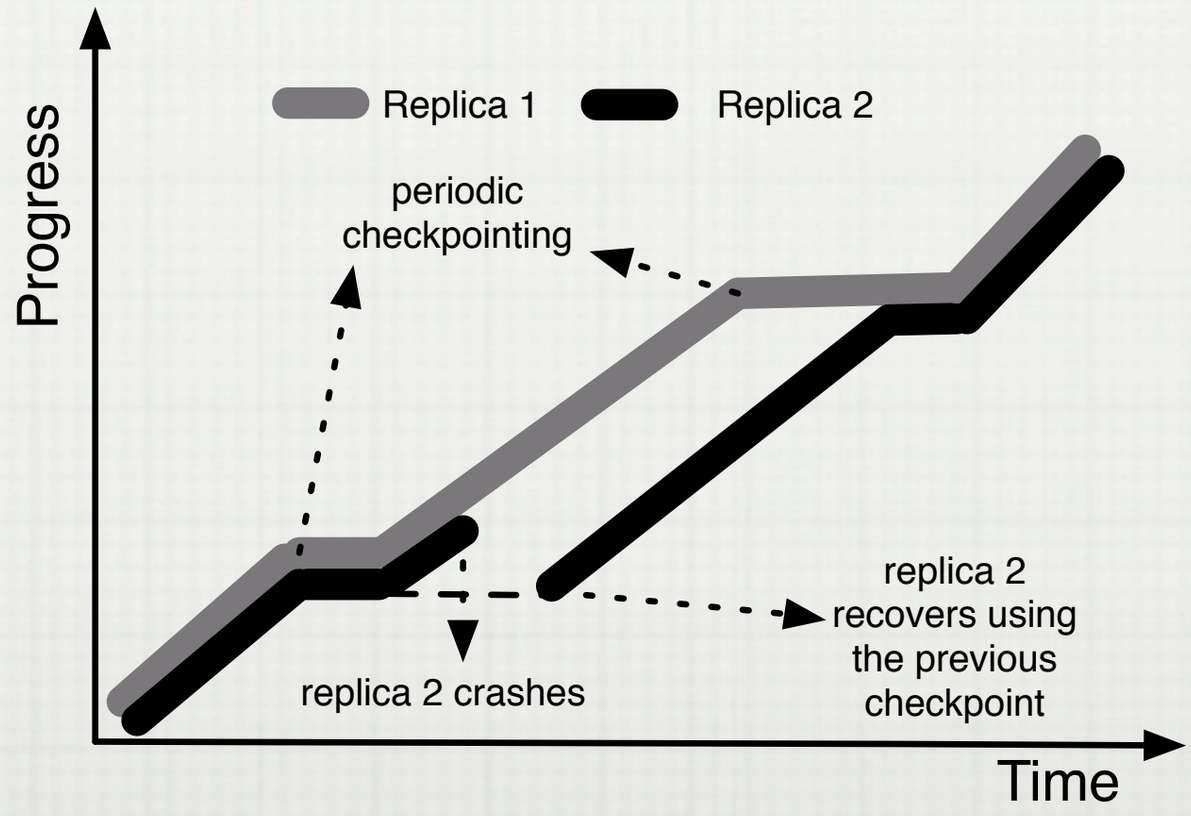
Comparison for soft data corruption

REPLICATION ENHANCED CHECKPOINTING

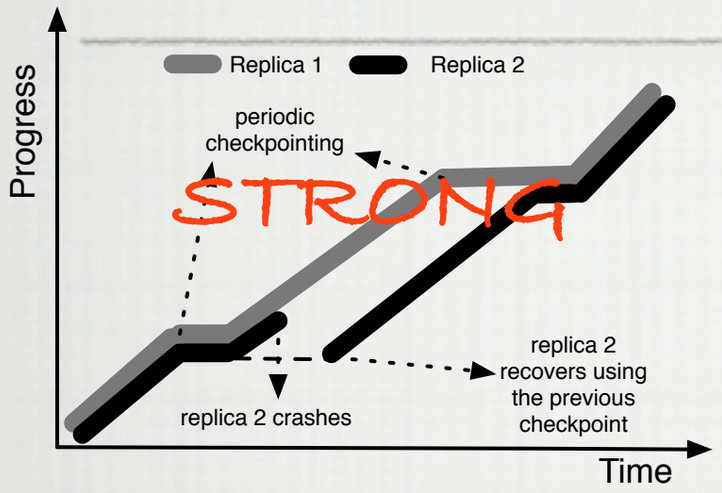


DIFFERENT WAYS TO RESTART FROM HARD ERROR

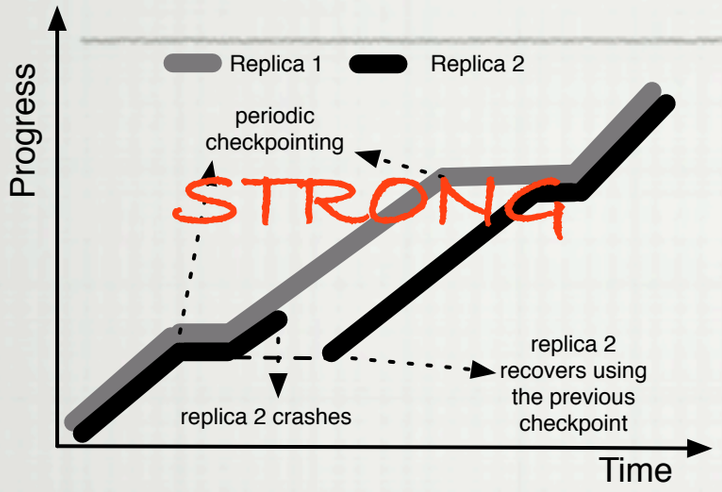
DIFFERENT WAYS TO RESTART FROM HARD ERROR

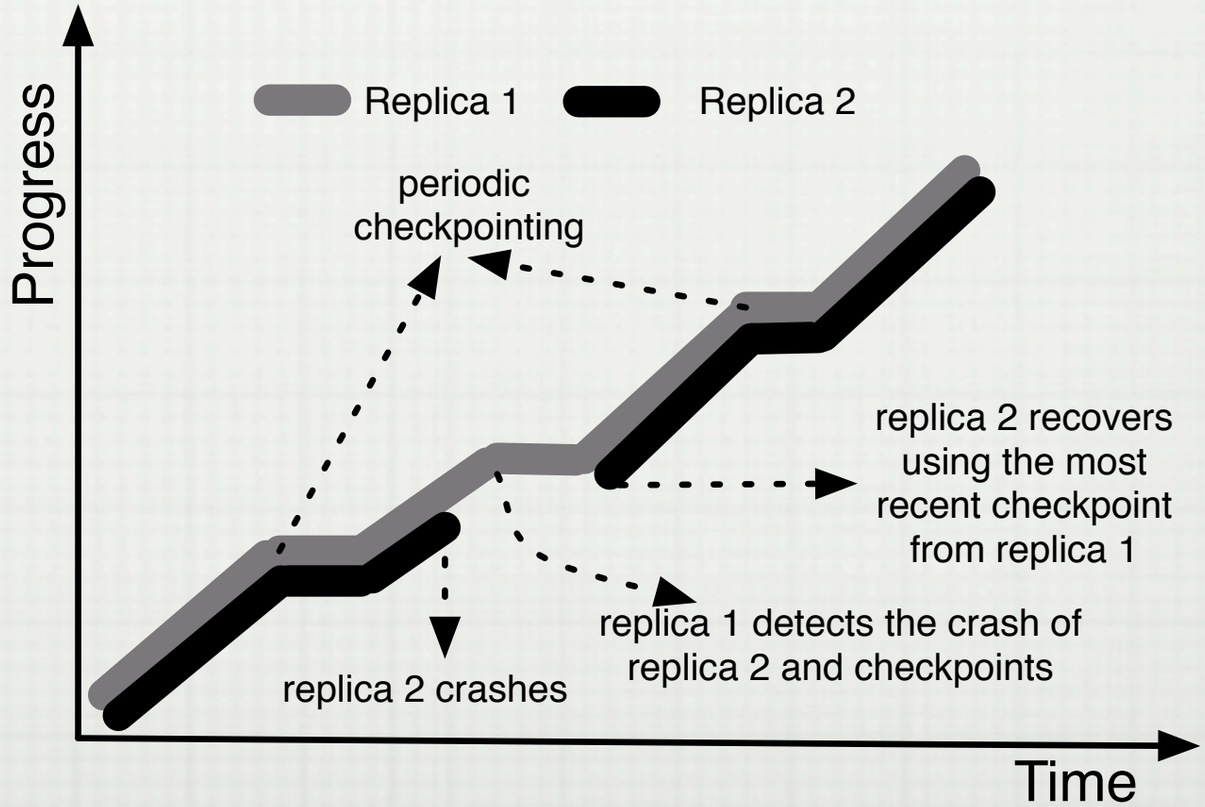


DIFFERENT WAYS TO RESTART FROM HARD ERROR

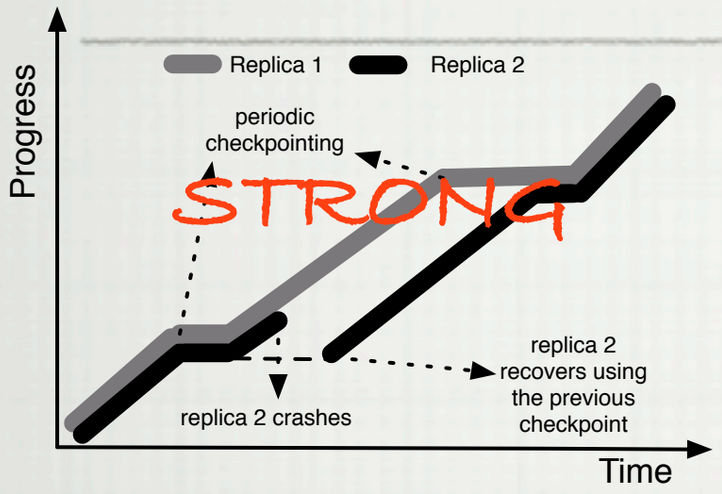


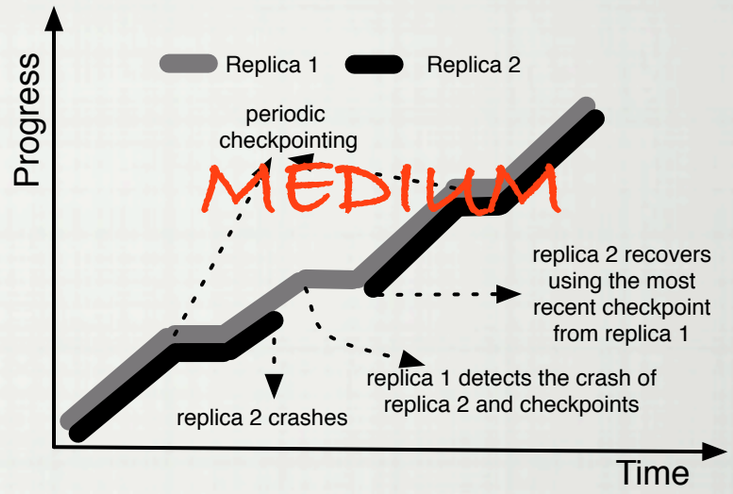
DIFFERENT WAYS TO RESTART FROM HARD ERROR



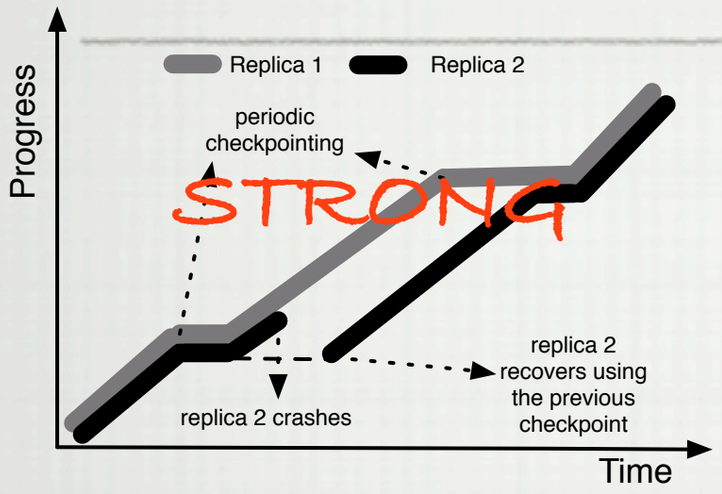


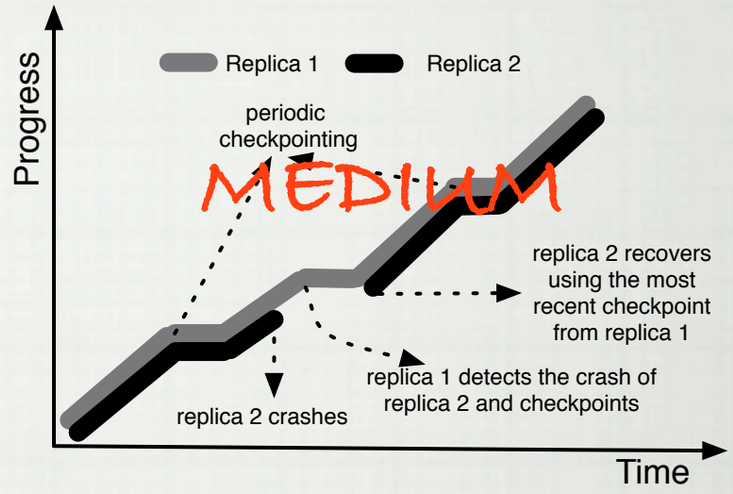
DIFFERENT WAYS TO RESTART FROM HARD ERROR

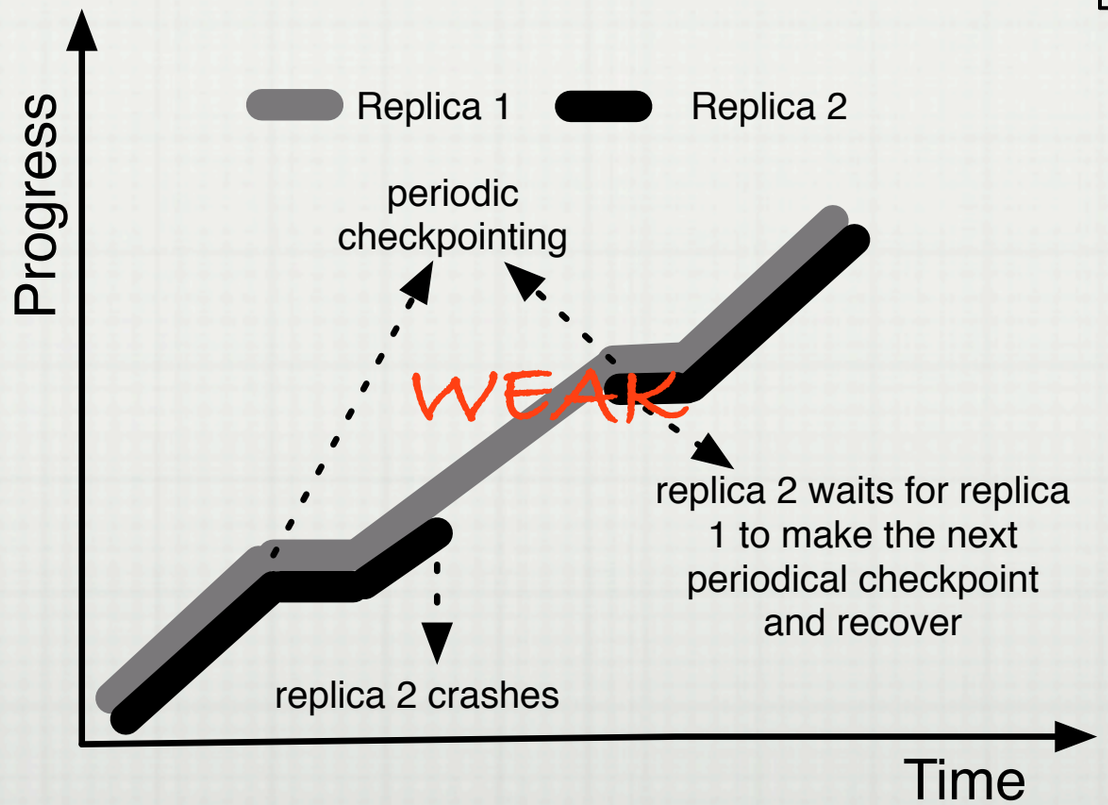




DIFFERENT WAYS TO RESTART FROM HARD ERROR



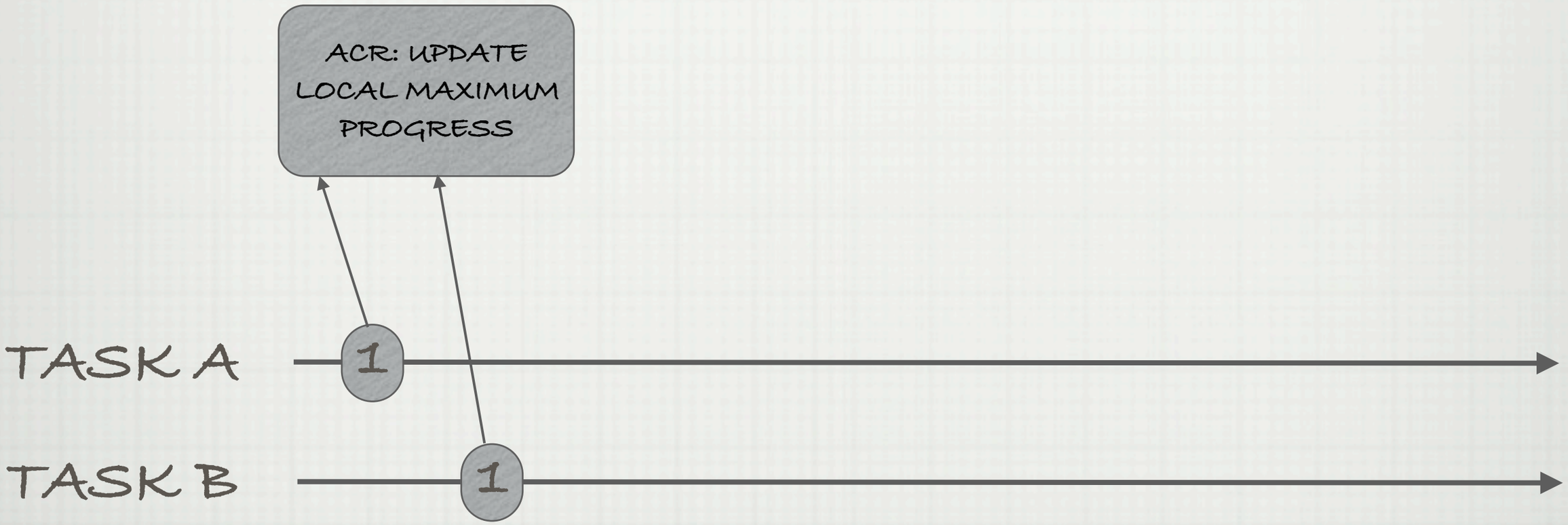




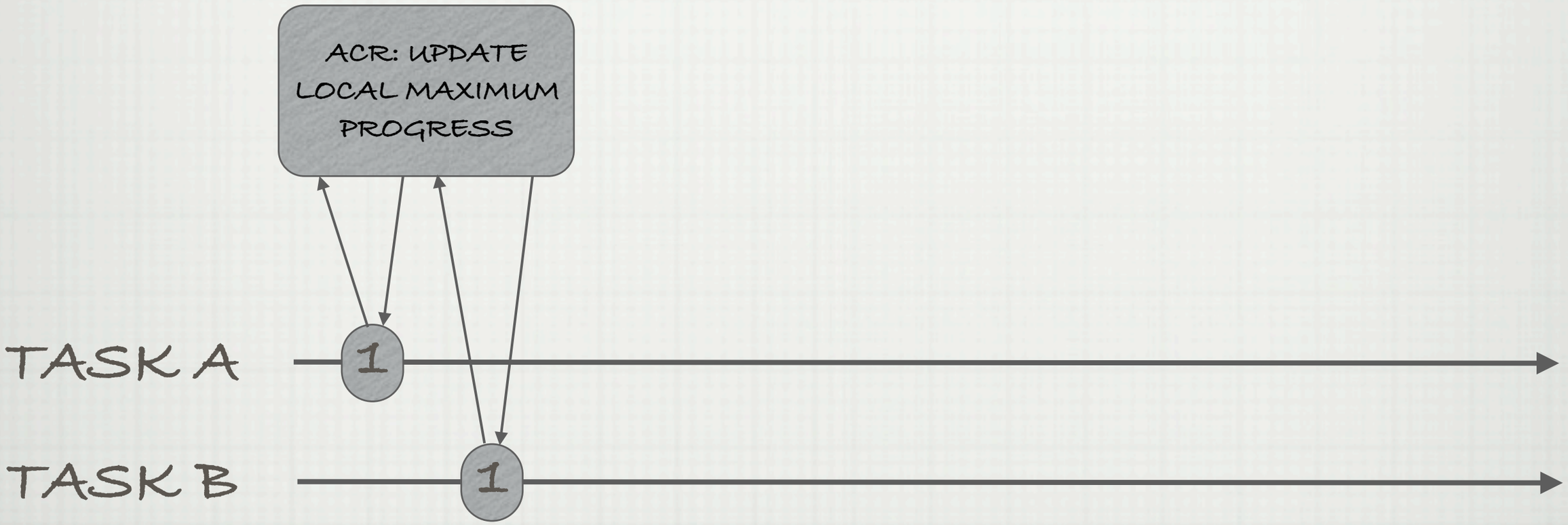
AUTOMATIC CHECKPOINT DECISION

ACR: UPDATE LOCAL MAXIMUM PROGRESS

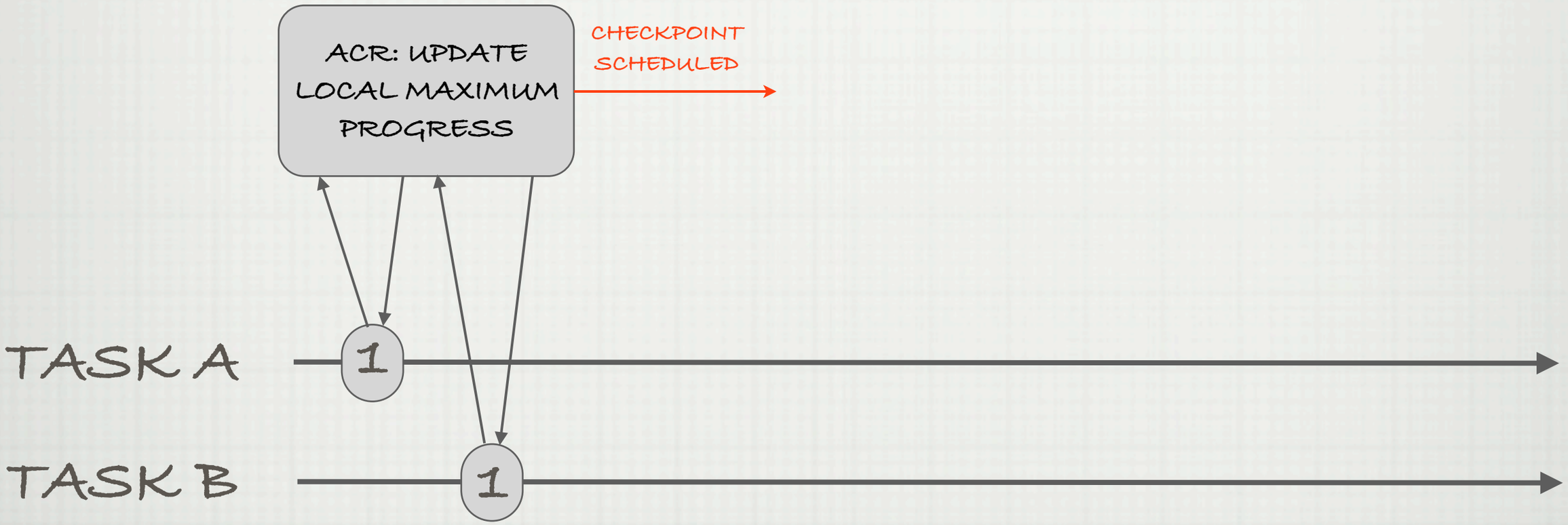
AUTOMATIC CHECKPOINT DECISION



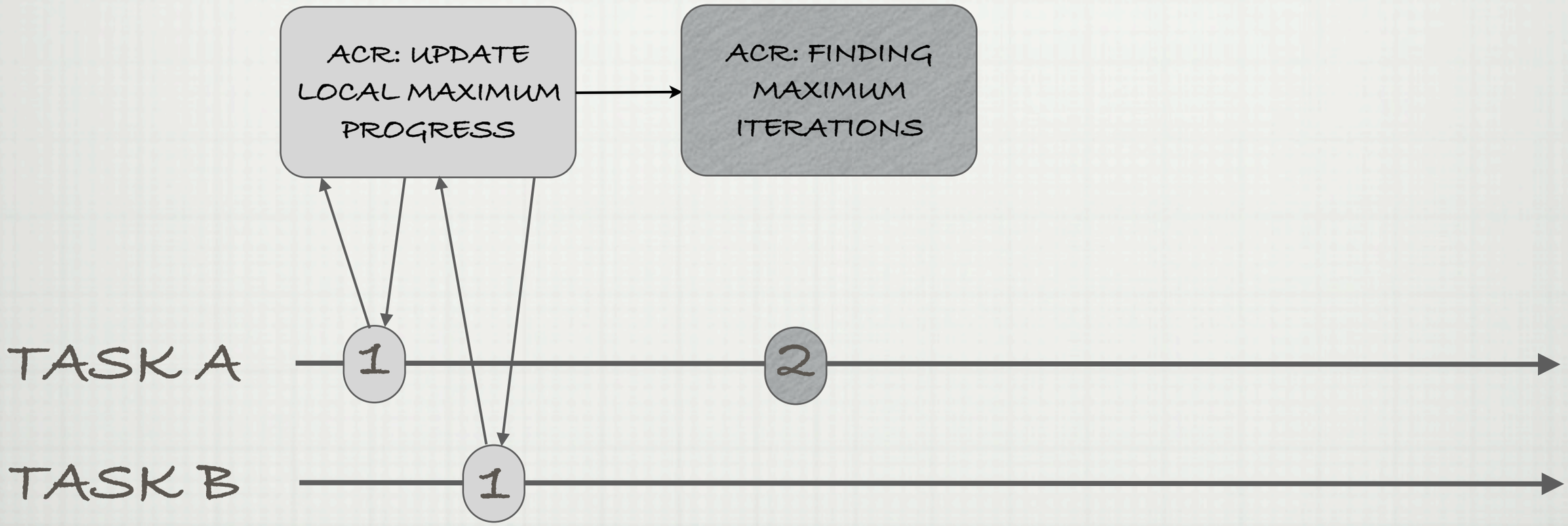
AUTOMATIC CHECKPOINT DECISION



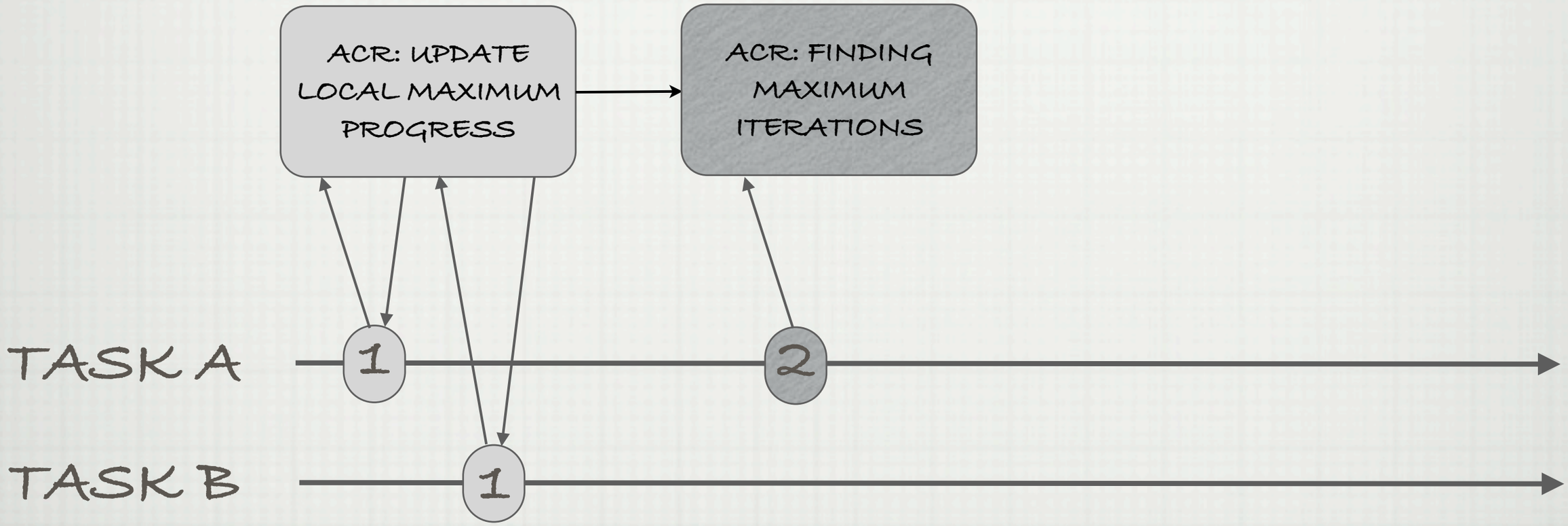
AUTOMATIC CHECKPOINT DECISION



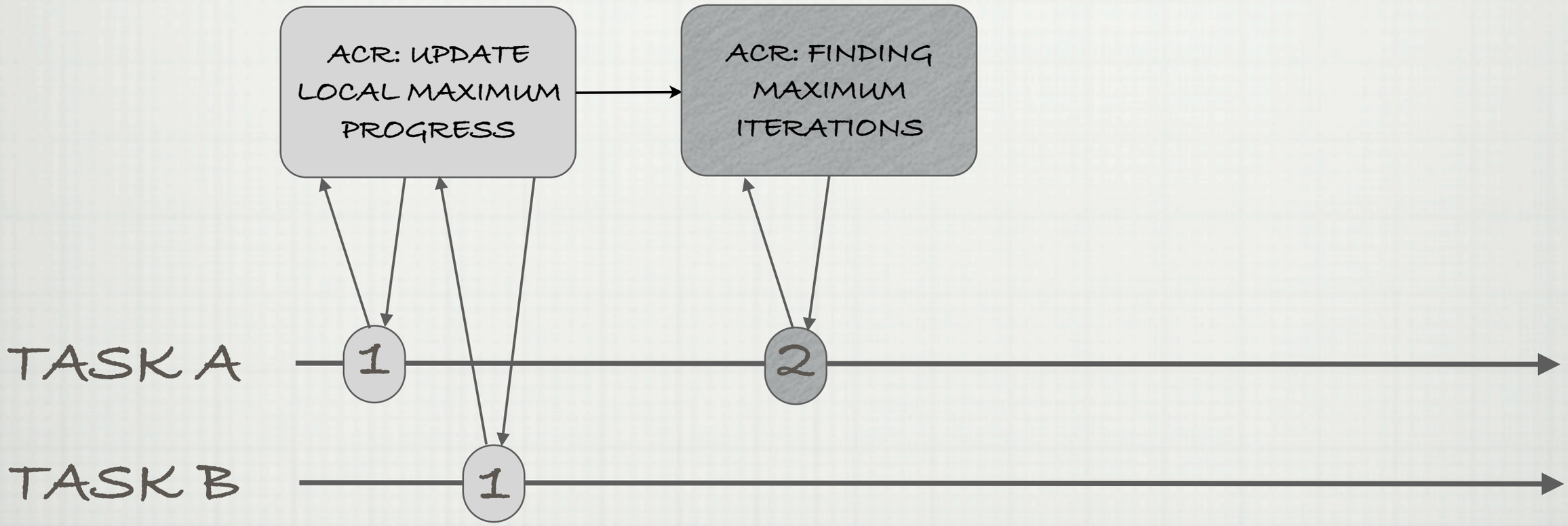
AUTOMATIC CHECKPOINT DECISION



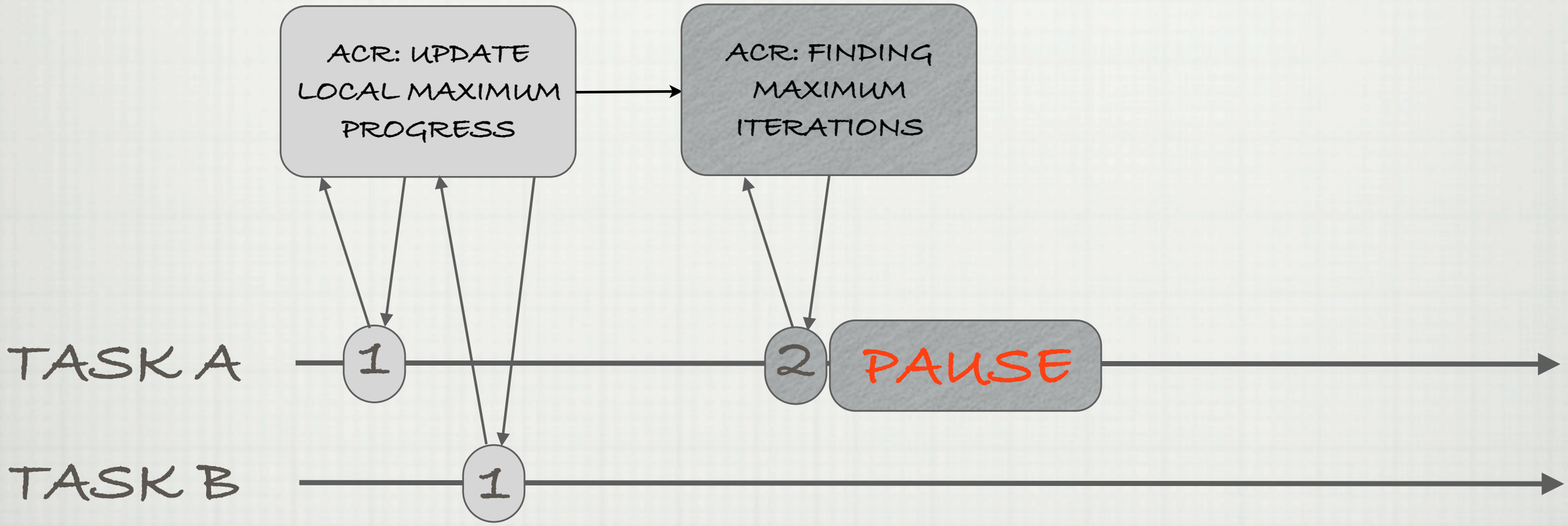
AUTOMATIC CHECKPOINT DECISION



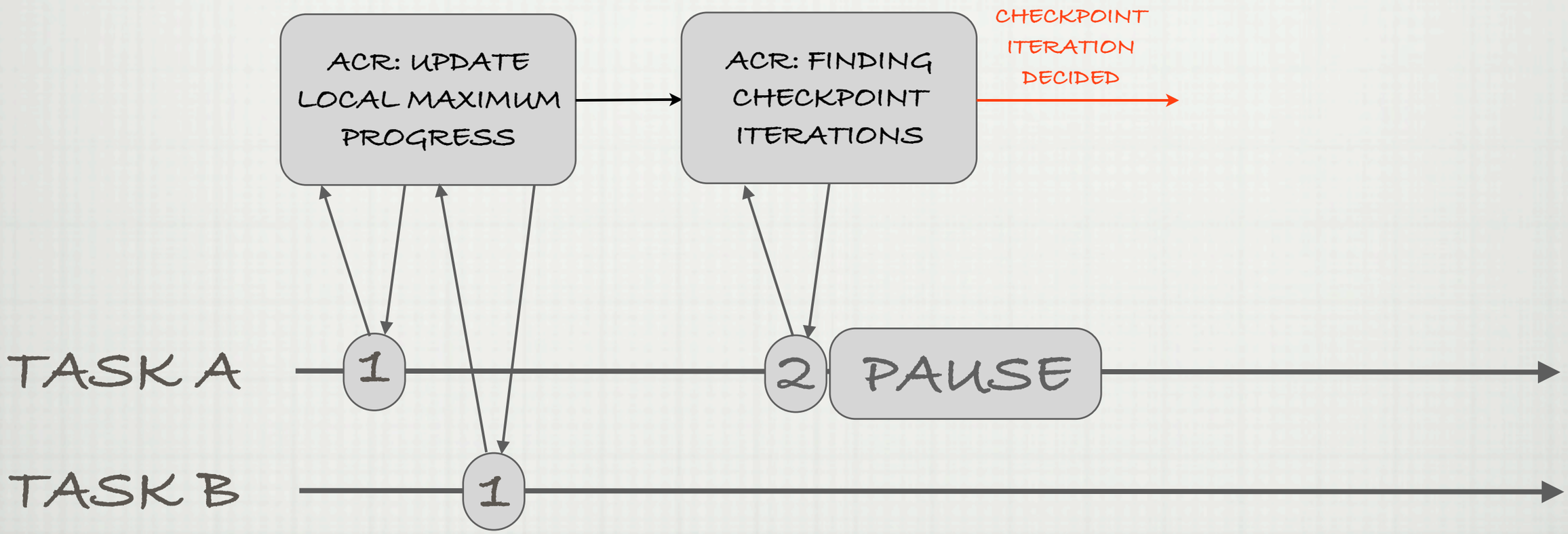
AUTOMATIC CHECKPOINT DECISION



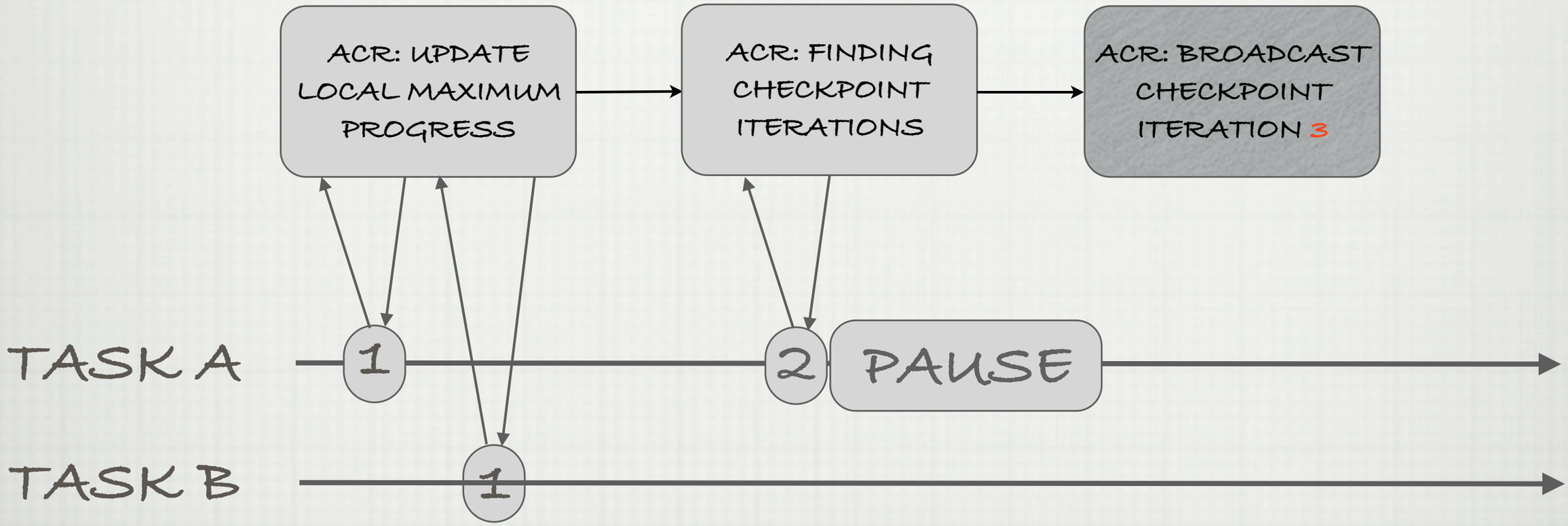
AUTOMATIC CHECKPOINT DECISION



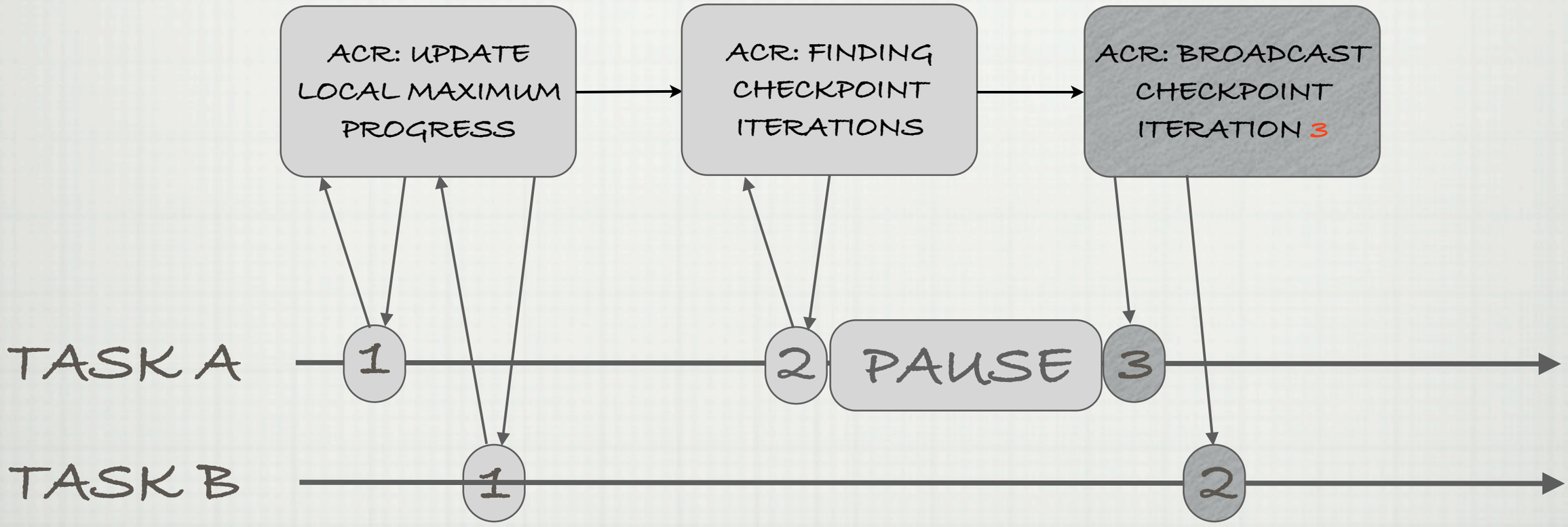
AUTOMATIC CHECKPOINT DECISION



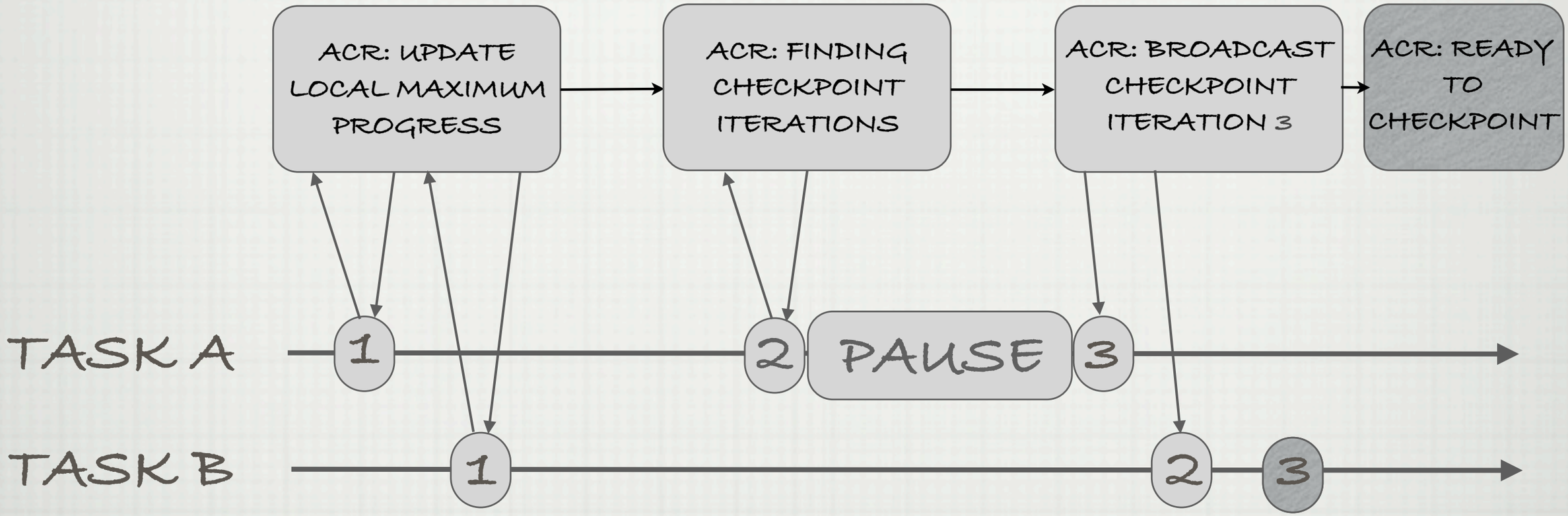
AUTOMATIC CHECKPOINT DECISION



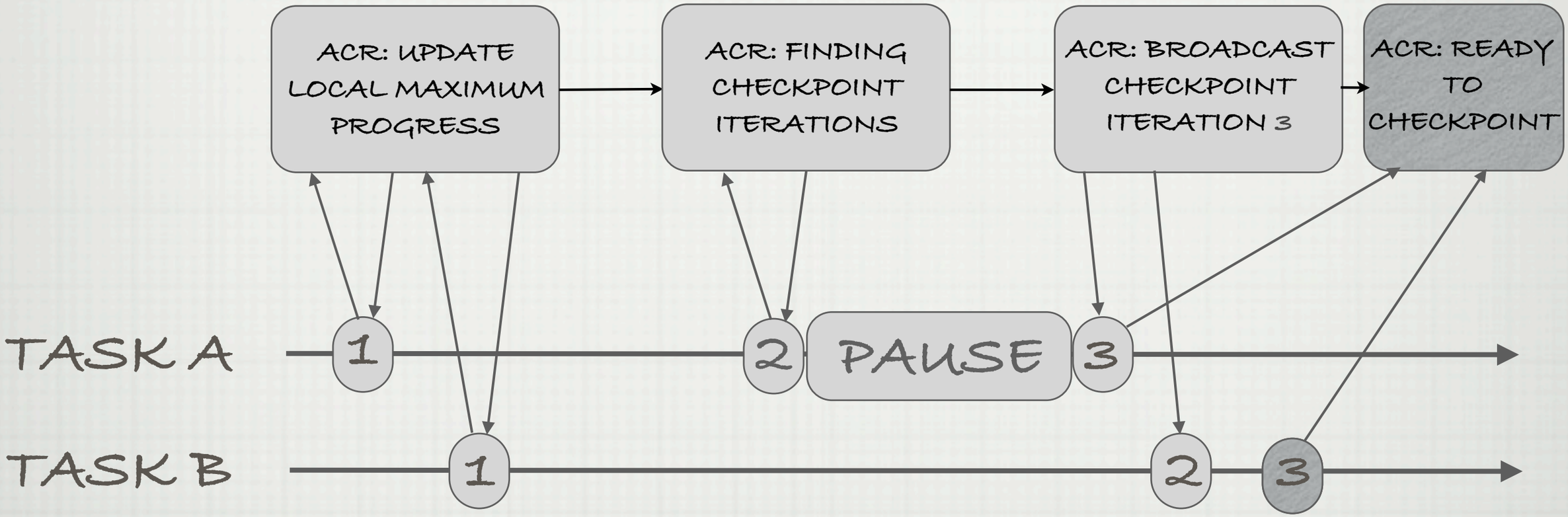
AUTOMATIC CHECKPOINT DECISION



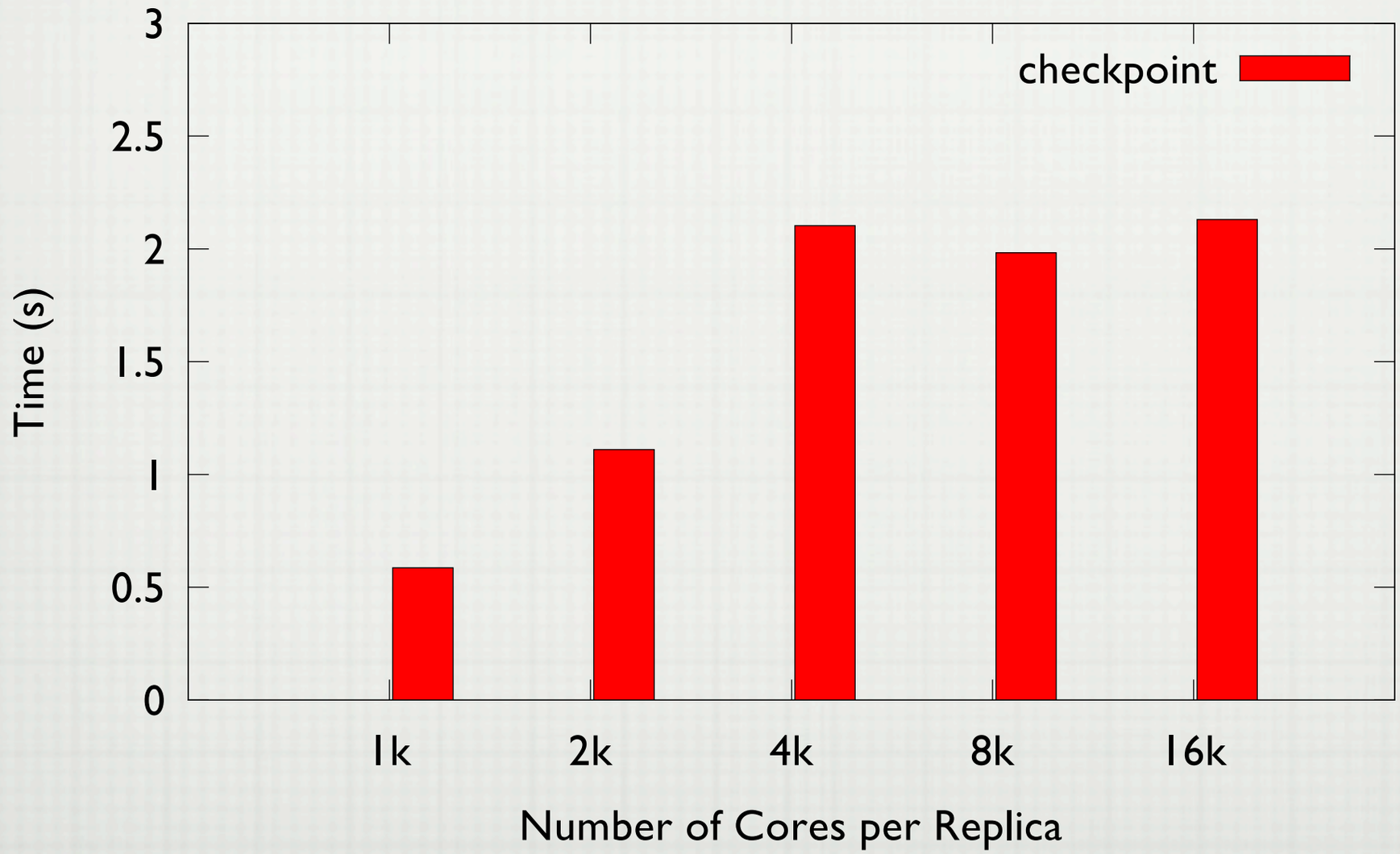
AUTOMATIC CHECKPOINT DECISION



AUTOMATIC CHECKPOINT DECISION

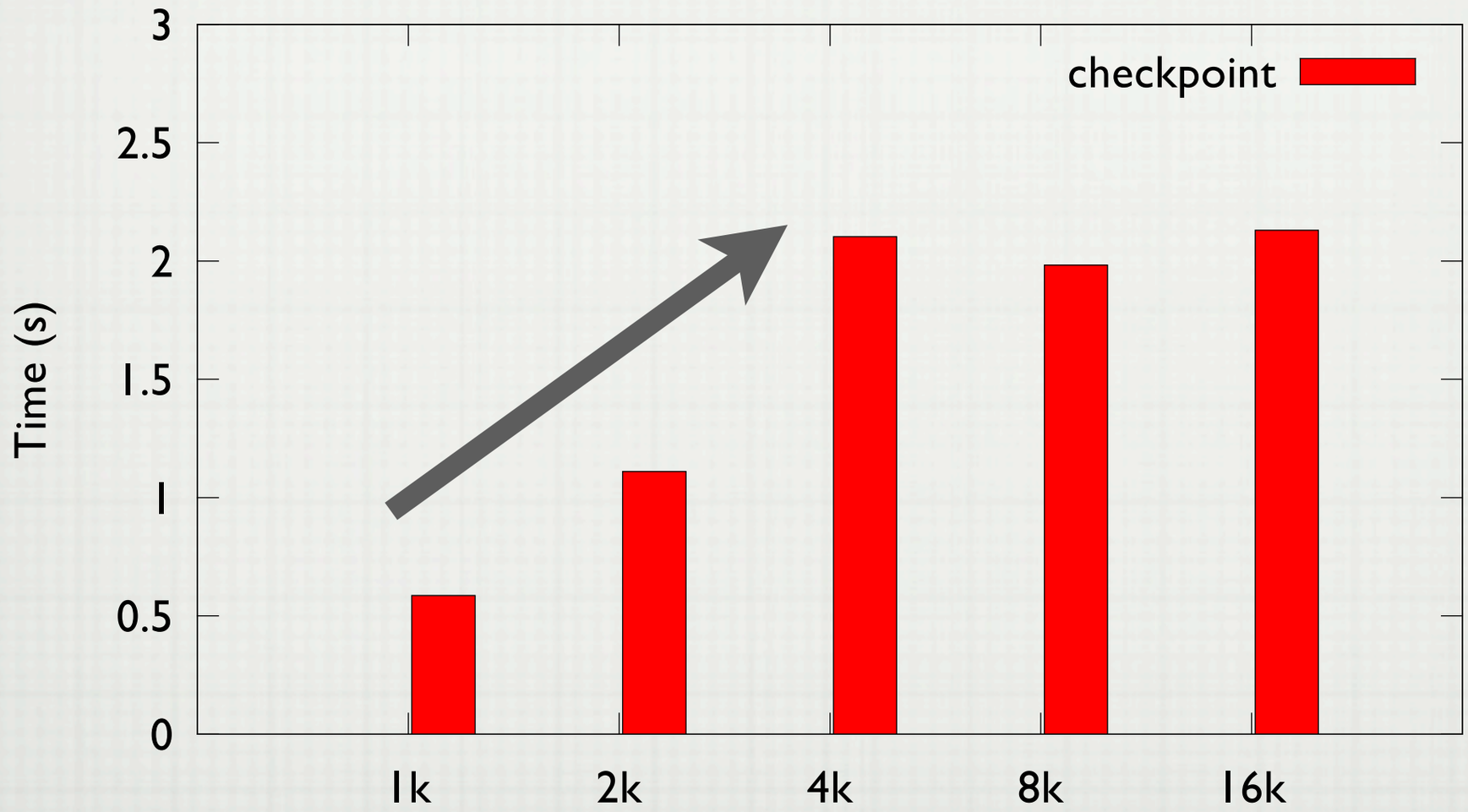


BASE PERFORMANCE



JACOBI3D BGP

BASE PERFORMANCE

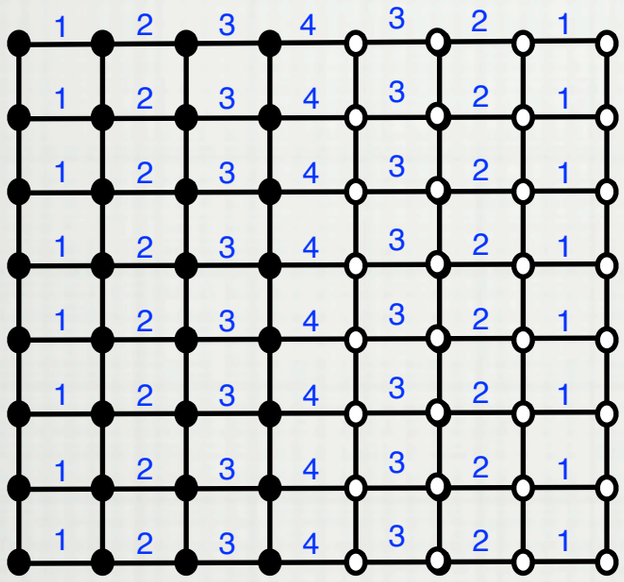


JACOBI3D BGP

OPTIMIZATION: TOPOLOGY AWARE MAPPING

OPTIMIZATION: TOPOLOGY AWARE MAPPING

- Replica 1 nodes



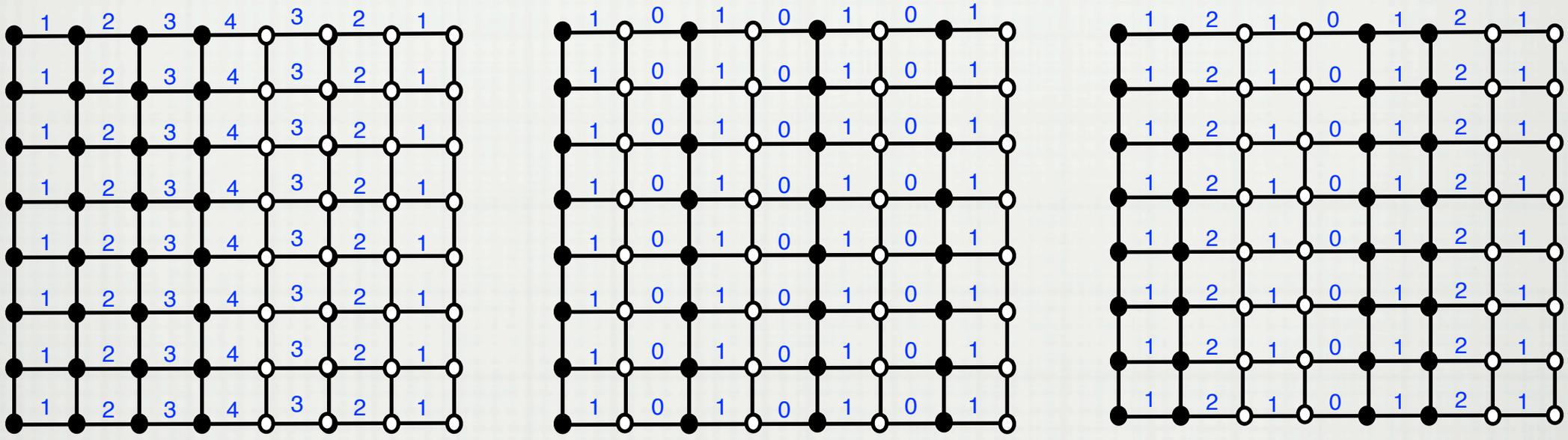
(a) Default-mapping

OPTIMIZATION: TOPOLOGY AWARE MAPPING

● Replica 1 nodes

o Replica 2 nodes

[0-4] # inter-replica messages



(a) Default-mapping

(b) Column-mapping

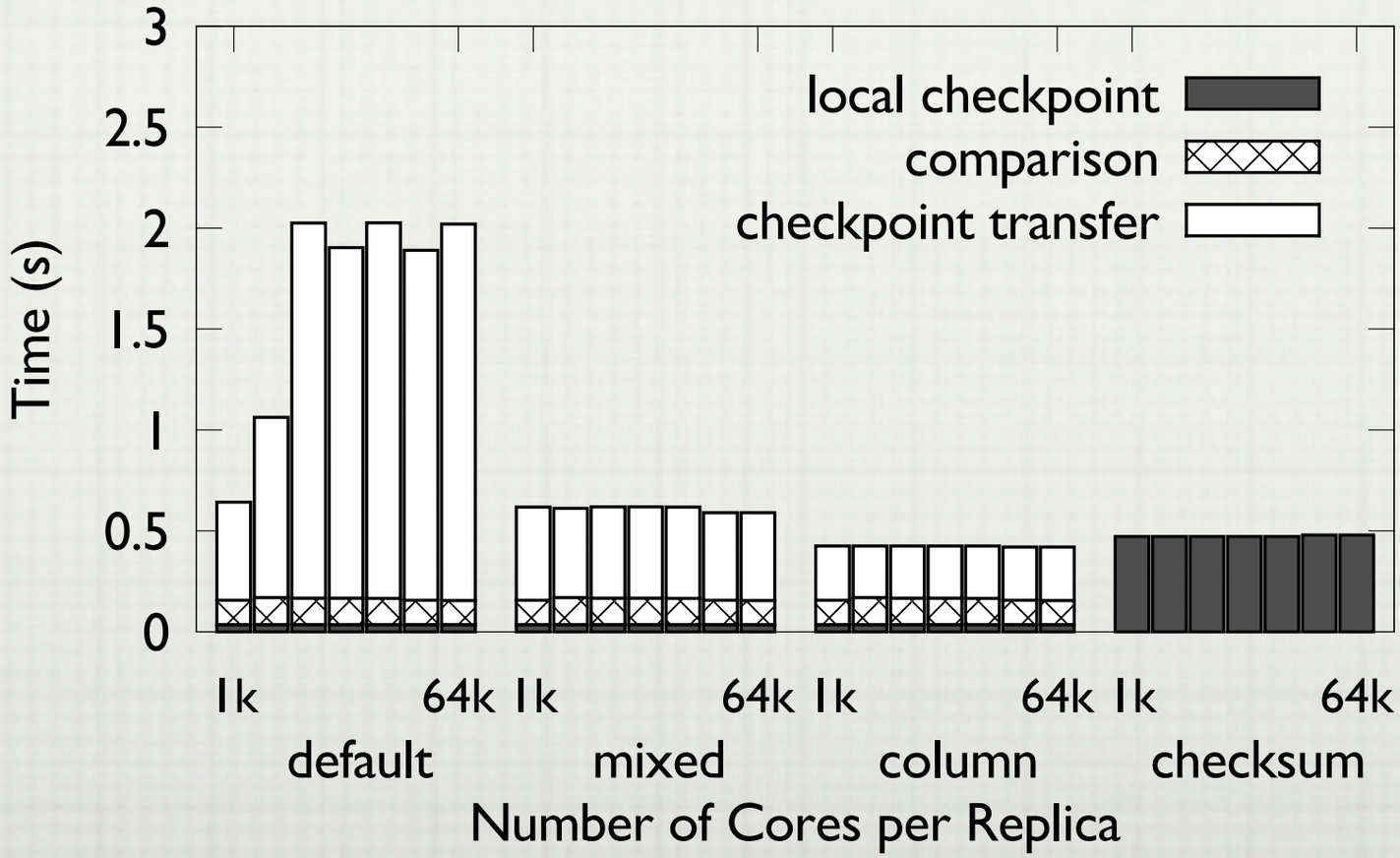
(c) Mixed-mapping

1) Reduce the inter-replica communication distance.

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

OPTIMIZATION: CHECKSUM

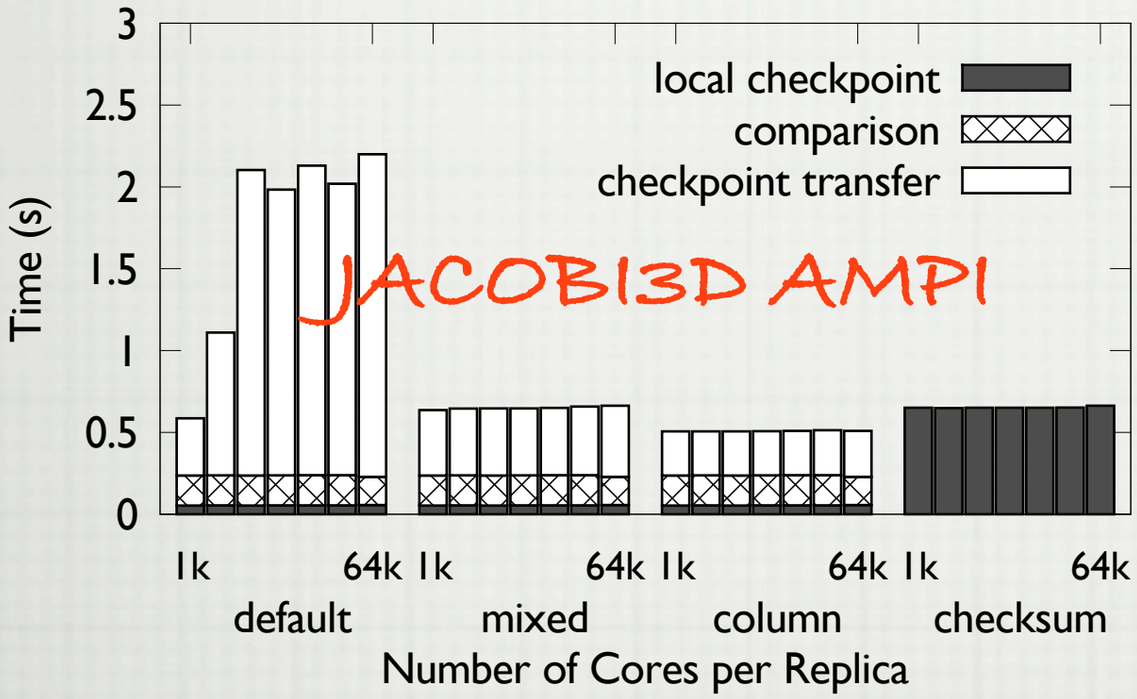
- ☐ TRANSFER THE CHECKSUM OF 1 INTEGER INSTEAD OF THE WHOLE CHECKPOINTS
- ☐ FLOATING POINT ROUND-OFF ERROR

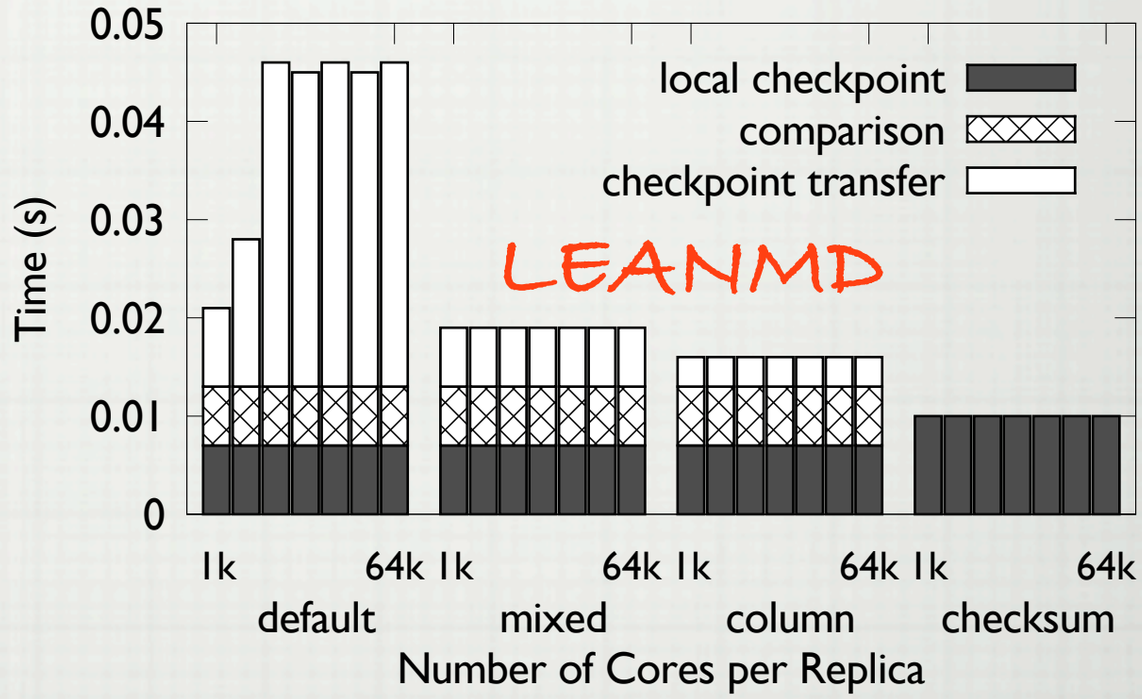


EXPERIMENTAL RESULTS: MINI-APPLICATIONS

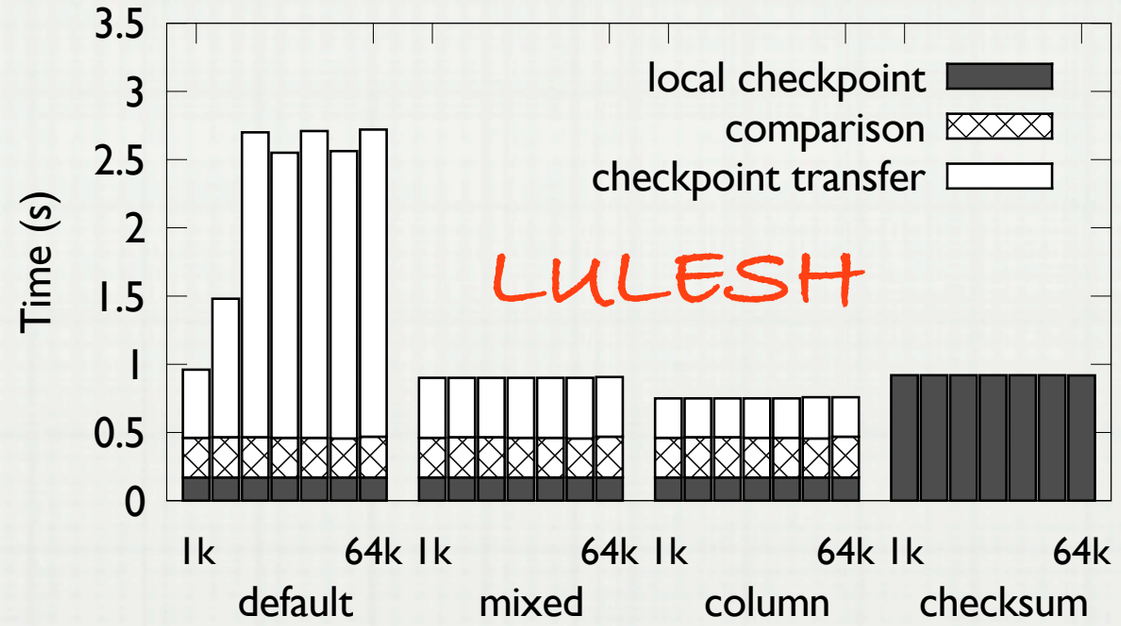
Benchmark	Description	Configuration per core	Memory Pressure
Jacobi3D	7-point stencil	64*64*128	High
HPCCG	Unstructured implicit finite element method	40*40*40	High
LULESH	Unstructured explicit mesh	32*32*64	High
LeanMD	Short-range non-bonded force calculation in NAMD	4000 atoms	Low
miniMD	Mimic the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

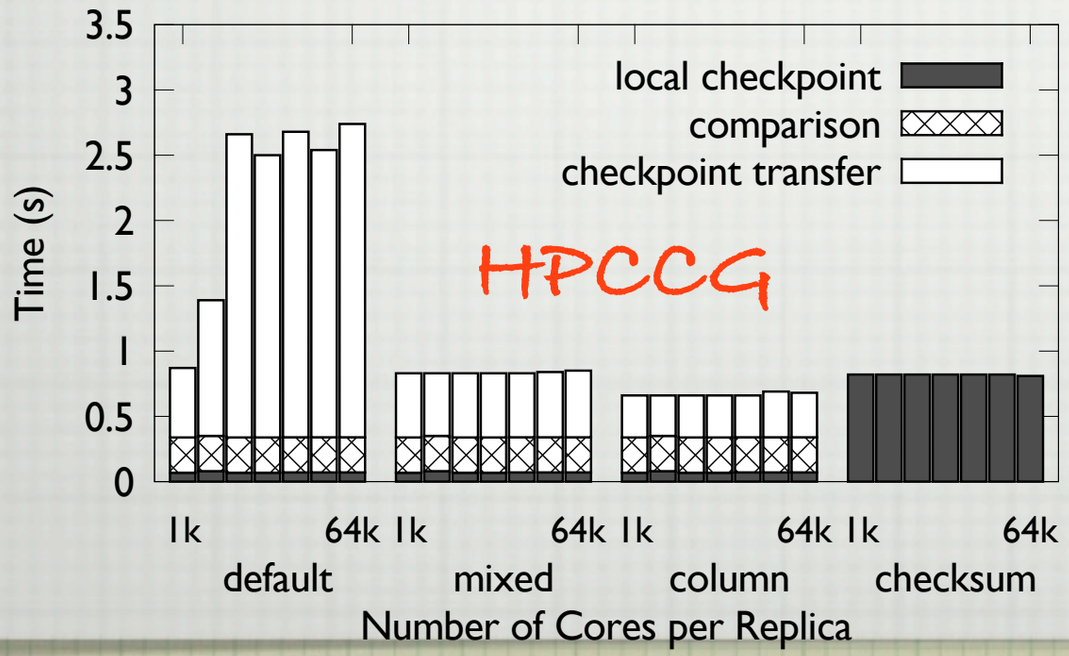
EXPERIMENTAL RESULTS: CHECKPOINT

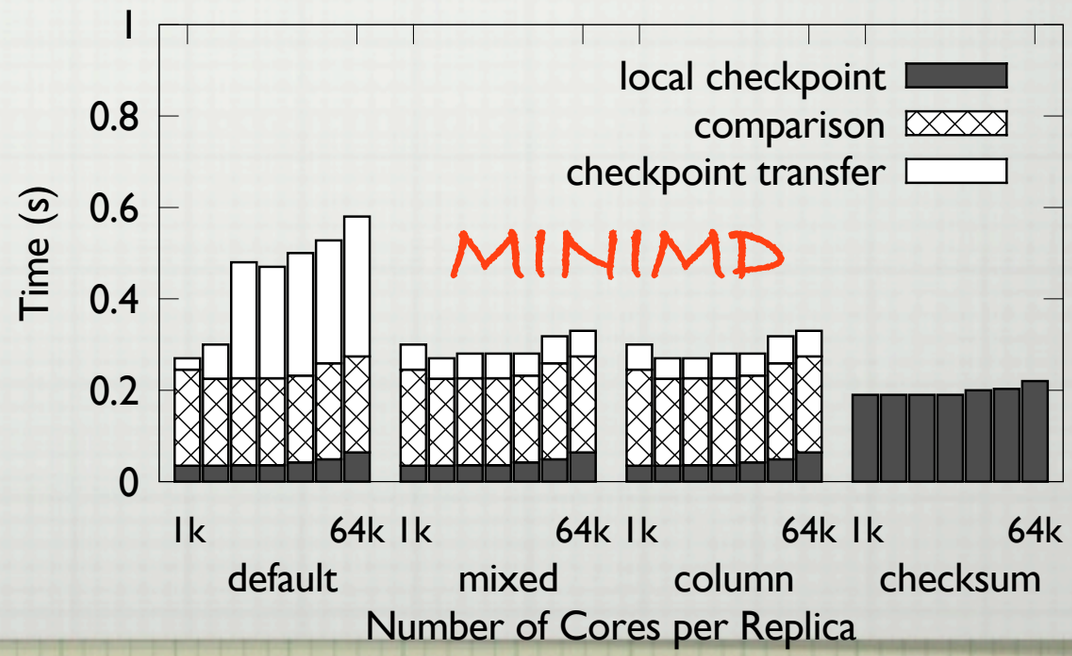




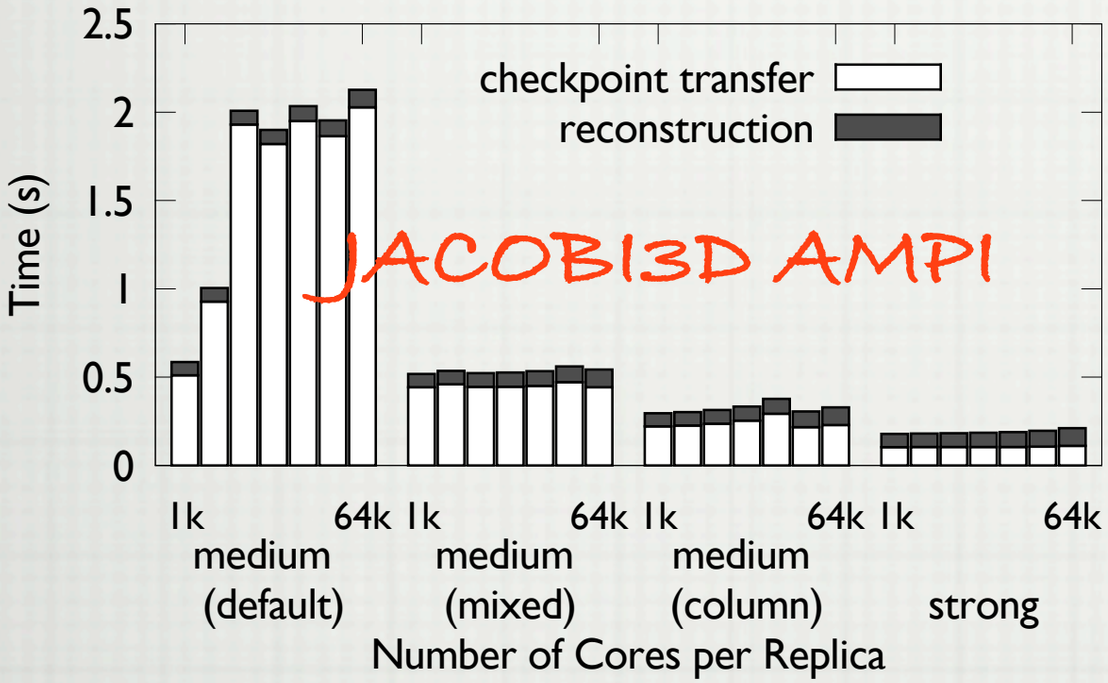
EXPERIMENTAL RESULTS: CHECKPOINT

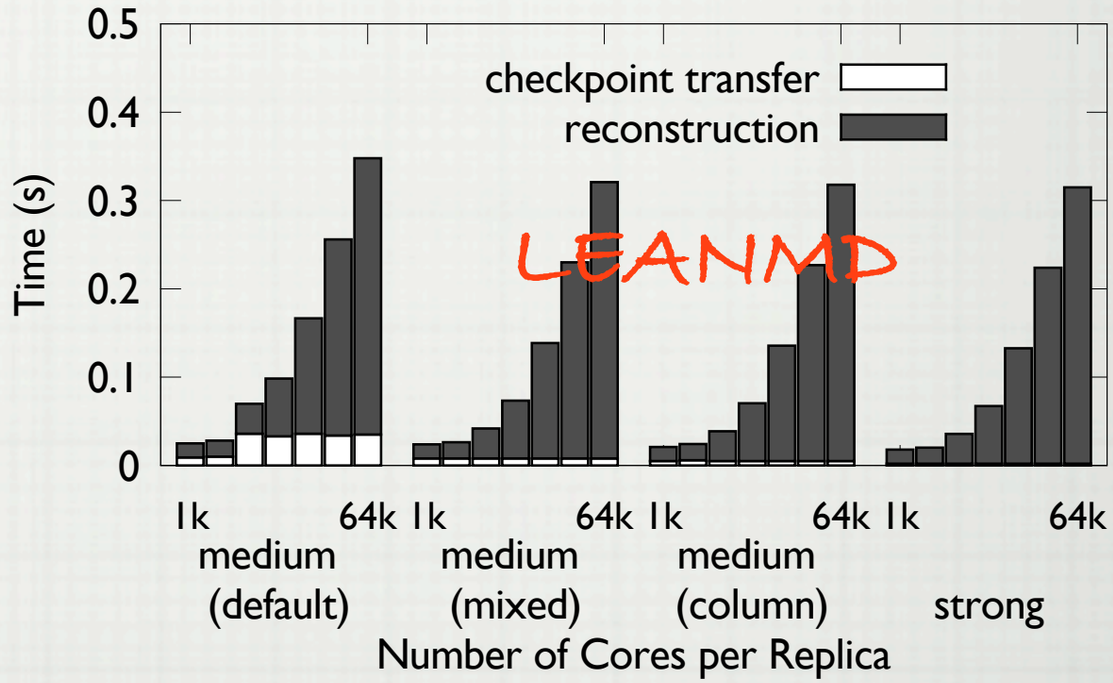




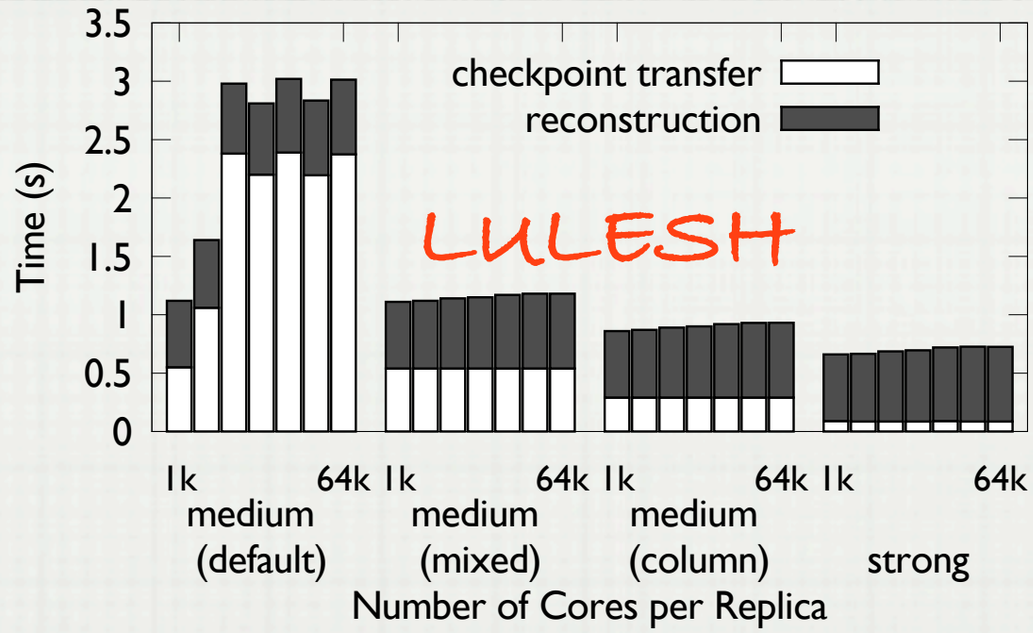


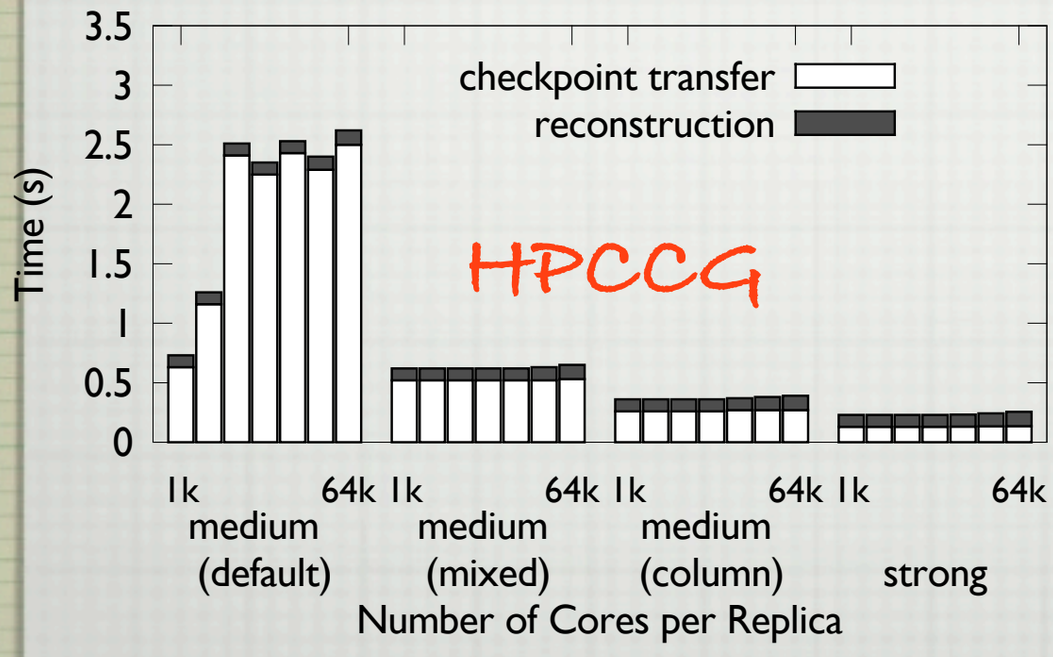
EXPERIMENTAL RESULTS: RESTART

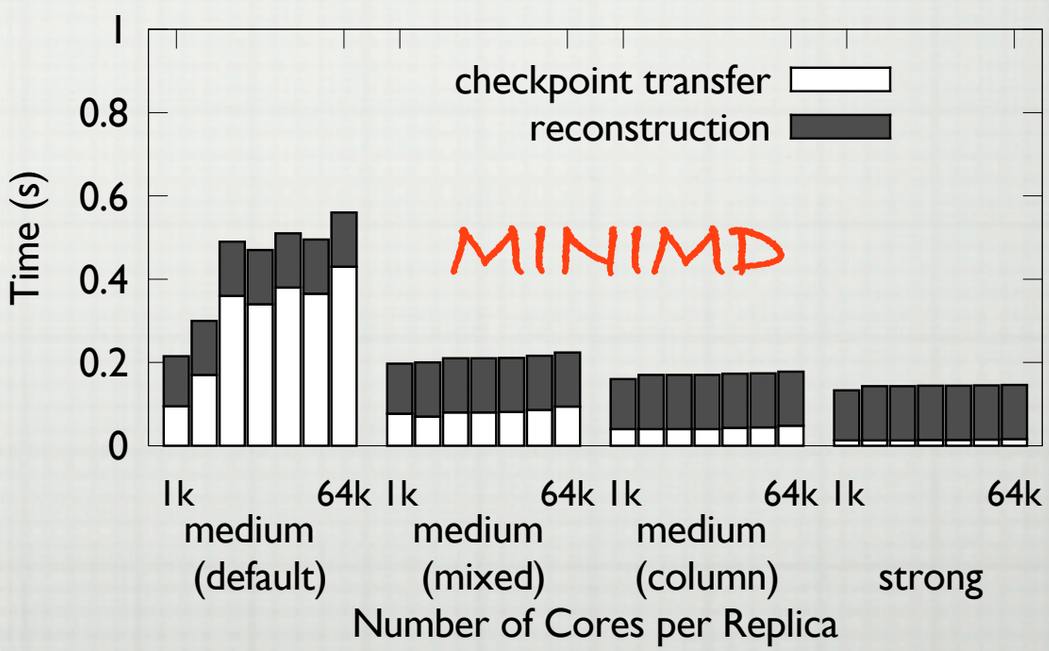




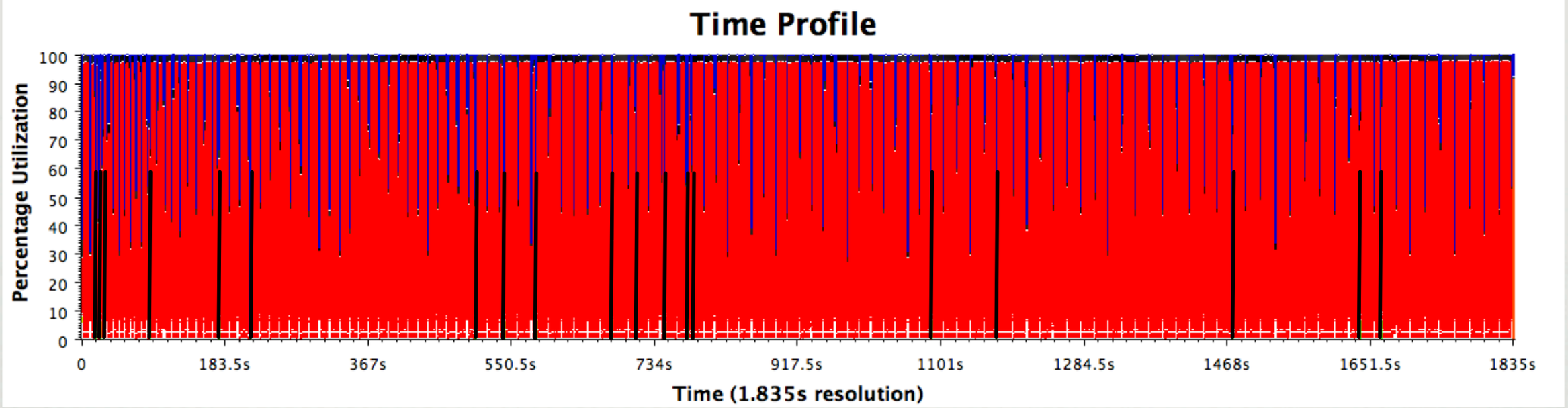
EXPERIMENTAL RESULTS: RESTART





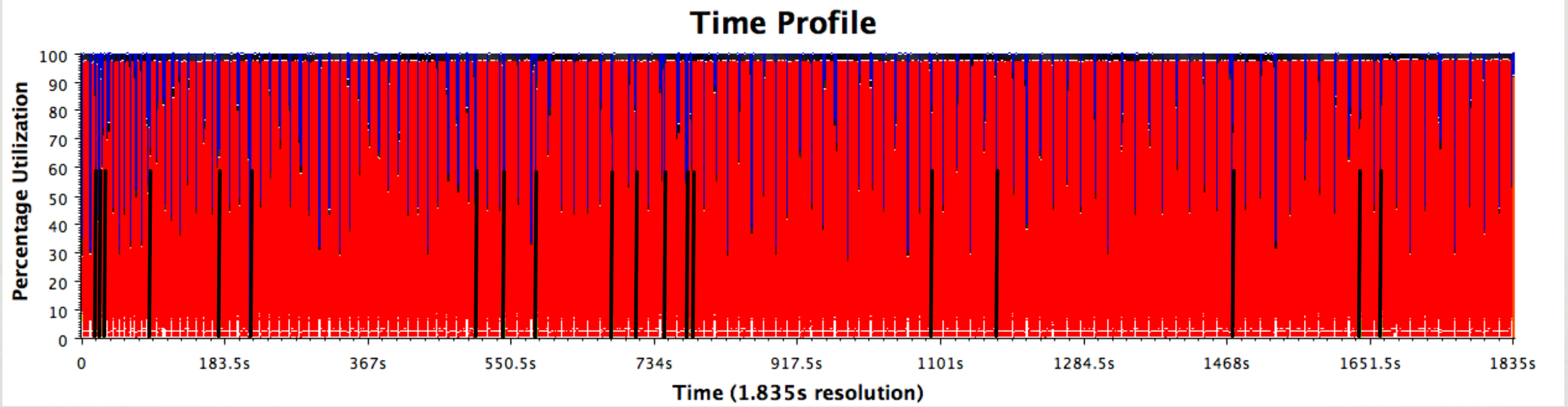


ADAPTING TO FAILURES



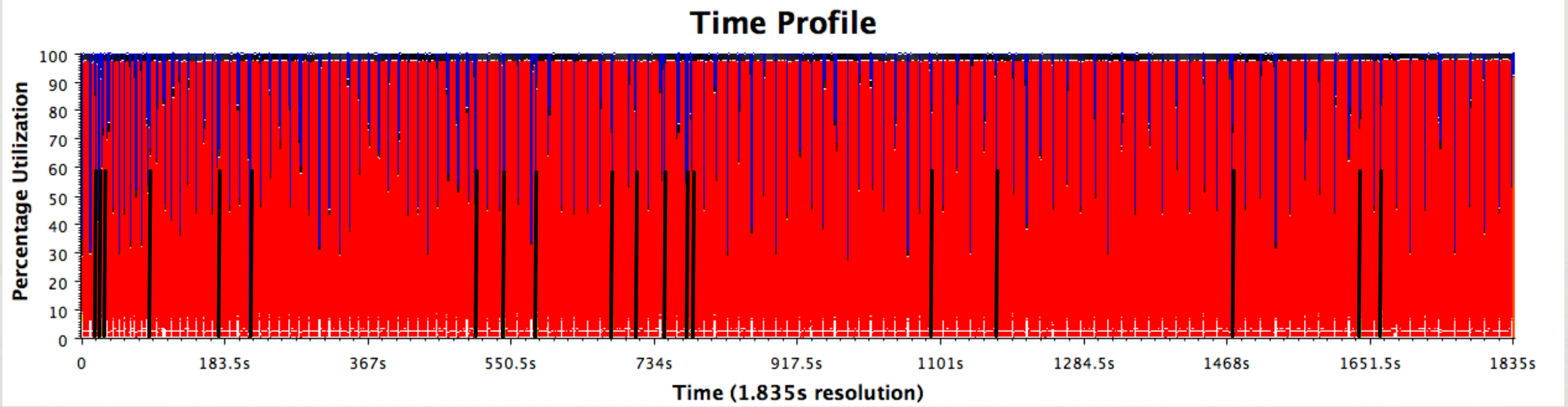
- Failures are injected according to Weibull process: shape parameter 0.6
- Changing Checkpoint period

ADAPTING TO FAILURES



- Failures are injected according to Weibull process: shape parameter 0.6
- Changing Checkpoint period

ADAPTING TO FAILURES

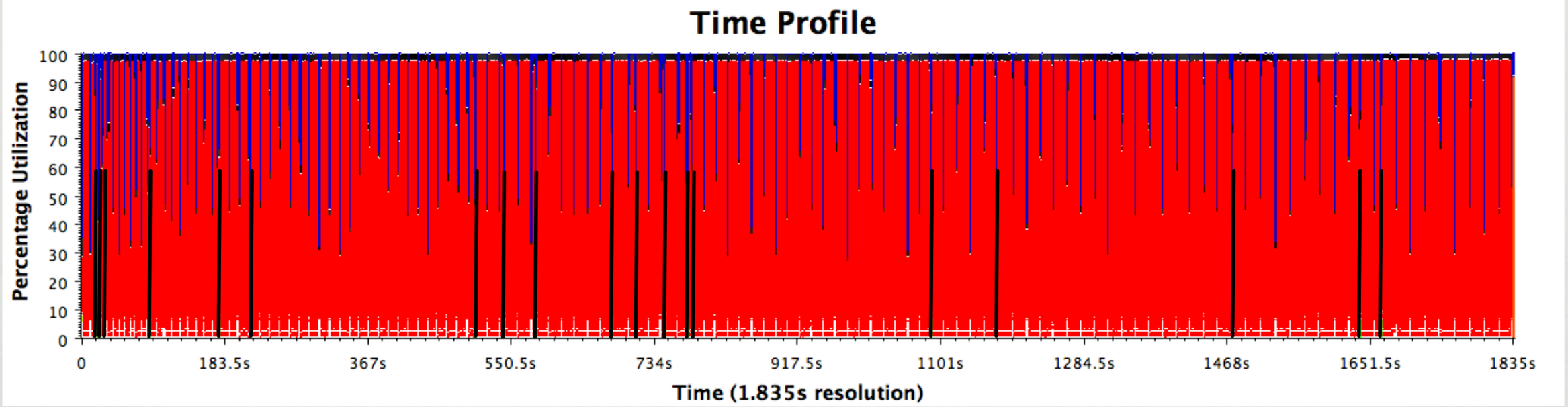


65

- Failures are injected according to Weibull process: shape parameter 0.6
- Changing Checkpoint period

175

ADAPTING TO FAILURES



65

- Failures are injected according to Weibull process: shape parameter 0.6
- Changing Checkpoint period

175

- Real failures injected: node becomes irresponsive
- Automatic restart: with the support of spare nodes and thus no need to submit the job again

CONCLUSION

- Automatic checkpoint decision
  - Restart from hard errors
  - Adapting to different failure distributions
- Protection for both hard and soft errors
- Good scalability

THANKS!

- [ ] QUESTIONS?