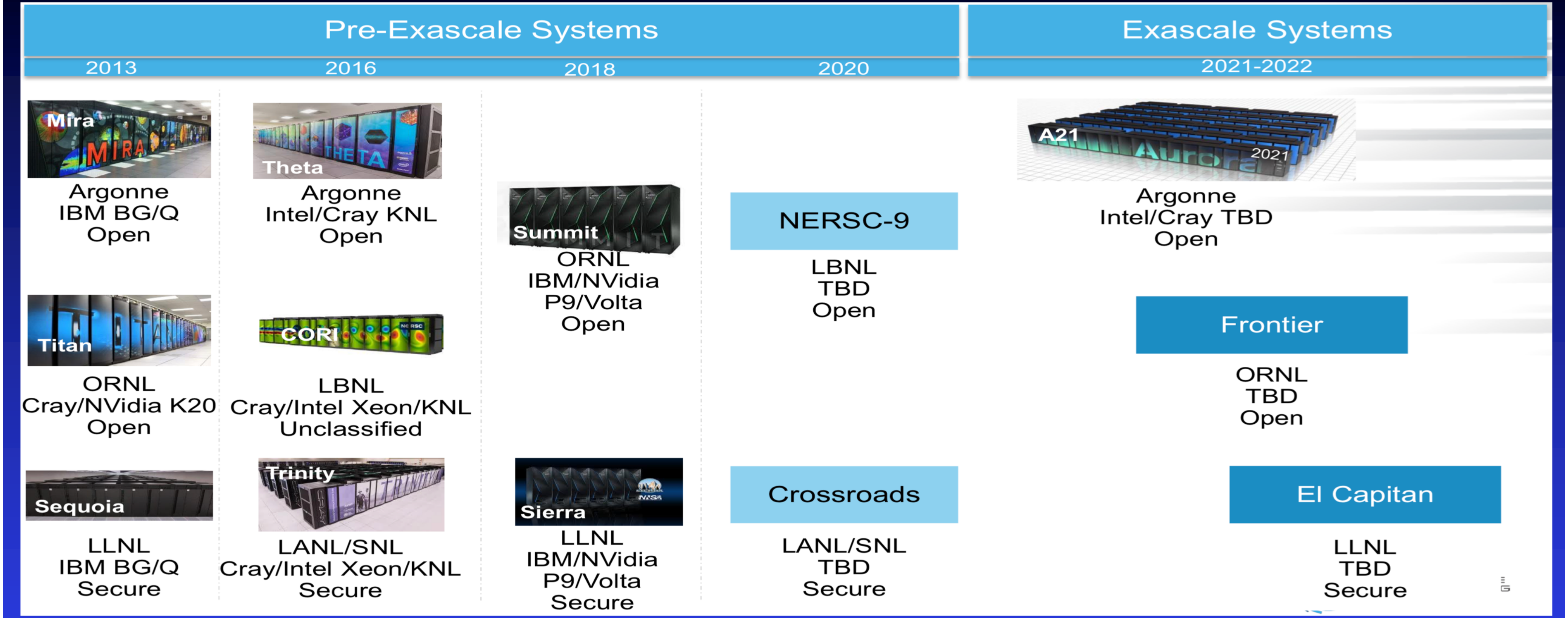
Evaluating Runtime and Power Requirements of Multilevel Checkpointing MPI Applications: An Empirical Study

Xingfu Wu, Valerie Taylor
Argonne National Laboratory, University of Chicago

Zhiling Lan
Illinois Institute of Technology

Motivations



Pre-Exascale Systems				Exascale Systems
2013	2016	2018	2020	2021-2022
Mira Argonne IBM BG/Q Open	Theta Argonne Intel/Cray KNL Open	Summit ORNL IBM/NVidia P9/Volta Open	NERSC-9 LBNL TBD Open	A21 Aurora 2021 Argonne Intel/Cray TBD Open
Titan ORNL Cray/NVidia K20 Open	CORI LBNL Cray/Intel Xeon/KNL Unclassified			Frontier ORNL TBD Open
Sequoia LLNL IBM BG/Q Secure	Trinity LANL/SNL Cray/Intel Xeon/KNL Secure	Sierra LLNL IBM/NVidia P9/Volta Secure	Crossroads LANL/SNL TBD Secure	El Capitan LLNL TBD Secure

- Real-world scientific applications often relies on fault tolerance techniques to successfully finish long executions because of faults in software and hardware
- Experiments are essential in order to fully understand how fault tolerant MPI applications impact both power and runtime on different architectures

Fault Tolerance Techniques

- **Checkpoint/restart** is a long-standing fault tolerance technique to alleviate the impact of system failures, in which an application save their state in a parallel file system because of a failure, then restart from the last saved checkpoint.
- **Redundancy approaches** improve resilience by replicating data or computation
- **Algorithm-based fault tolerance** maintains a coded global consistent state of the computation in memory by modifying applications to operate on encoded data
- **Proactive methods** take preventive actions before failures, such as process or object migration

FTI (Fault Tolerance Interface)

- FTI is a middleware library that offers multiple fault tolerance features through an easy-to-use interface to enhance the reliability of supercomputers
- FTI is written in C, and it targets high performance computing applications using MPI
- Lead developer: Leonardo Bautista Gomez from ANL (BSC now)
- Default four-level checkpointing configuration is ckp(3,5,7,11):
  - 3 minutes for L1, 5 minutes for L2,
  - 7 minutes for L3, and 11 minutes for L4.

FTI: Four level Checkpointing

- Level 1: Local Storage
  - Fastest checkpoint level, low reliability
  - No hardware failure (software failure)
- Level 2: Partner Copy
  - Checkpoint replication (Copy to neighbor node)
  - Tolerates single node crash
- Level 3: RS (Reed-Solomon) encoding
  - Checkpoint encoding
  - Tolerates multiple node crashes
- Level 4: Parallel File System
  - Classic checkpoint (the slowest level, largest output)
  - The most reliable level, tolerates power outage

Four Architectures

System Name	**Cray XC40 Theta**	**IBM BG/Q Mira**	**Linux Cluster Shepard**	**Linux Cluster Cooper**
Architecture	**Intel KNL**	**IBM BG/Q**	**Intel Haswell**	**AMD Kaveri**
Number of nodes	3,624	49,152	36	36
CPU cores per node	64	16	32	4
Sockets per node	1	1	2	1
CPU type and speed	Xeon Phi KNL 7230 1.30GHz	PowerPC A2 1.6GHz	Xeon(R) E5-2698 V3 2.3GHz	AMD A10-7850K 3.7GHz
L1 cache per core	D:32KB/I:32KB	D:16KB/I:16KB	D:32KB/I:32KB	D:16KB/I:96KB
L2 cache per socket	32MB (shared)	32MB (shared)	256KB (per core)	2MB (shared)
L3 cache per socket	None	None	40MB (shared)	None
Memory per node	16GB/192GB	16GB	128GB	16GB
Network	Cray Aries Dragonfly	5D Torus	Mellanox FDR InfiniBand	Mellanox FDR InfiniBand
Power tools	CapMC/PoLiMEr	EMON/MonEQ	PowerInsight	PowerInsight
TDP per socket	215W	55W	135W	65W
Power Management	Yes	No	No	Yes
File System	Lustre PFS	GPFS	Regular NFS	Regular NFS

Our Approaches

- Communication-intensive: We developed an FTI version of Intel MPI Benchmarks (IMB) and used it with the default checkpointing configuration to quantify the overhead of FTI
- Compute-intensive: We used the Heat Distribution Code (HDC) to investigate the performance and power impacts under different FTI configurations
- Memory-intensive: We used the memory benchmark STREAM to investigate the performance and power impacts under different FTI configurations

Intel MPI Benchmarks (IMB) with FTI

- IMB performs a set of MPI performance measurements for point-to-point and global communication operations for a range of message sizes (default from 1 byte to 4 MB)
- We developed an FTI version of IMB and used it with the default checkpointing configuration to quantify the overhead of FTI
- Overall, our experimental results show that the overhead of FTI is less than 10% in most cases

Heat Distribution Code (HDC) with FTI

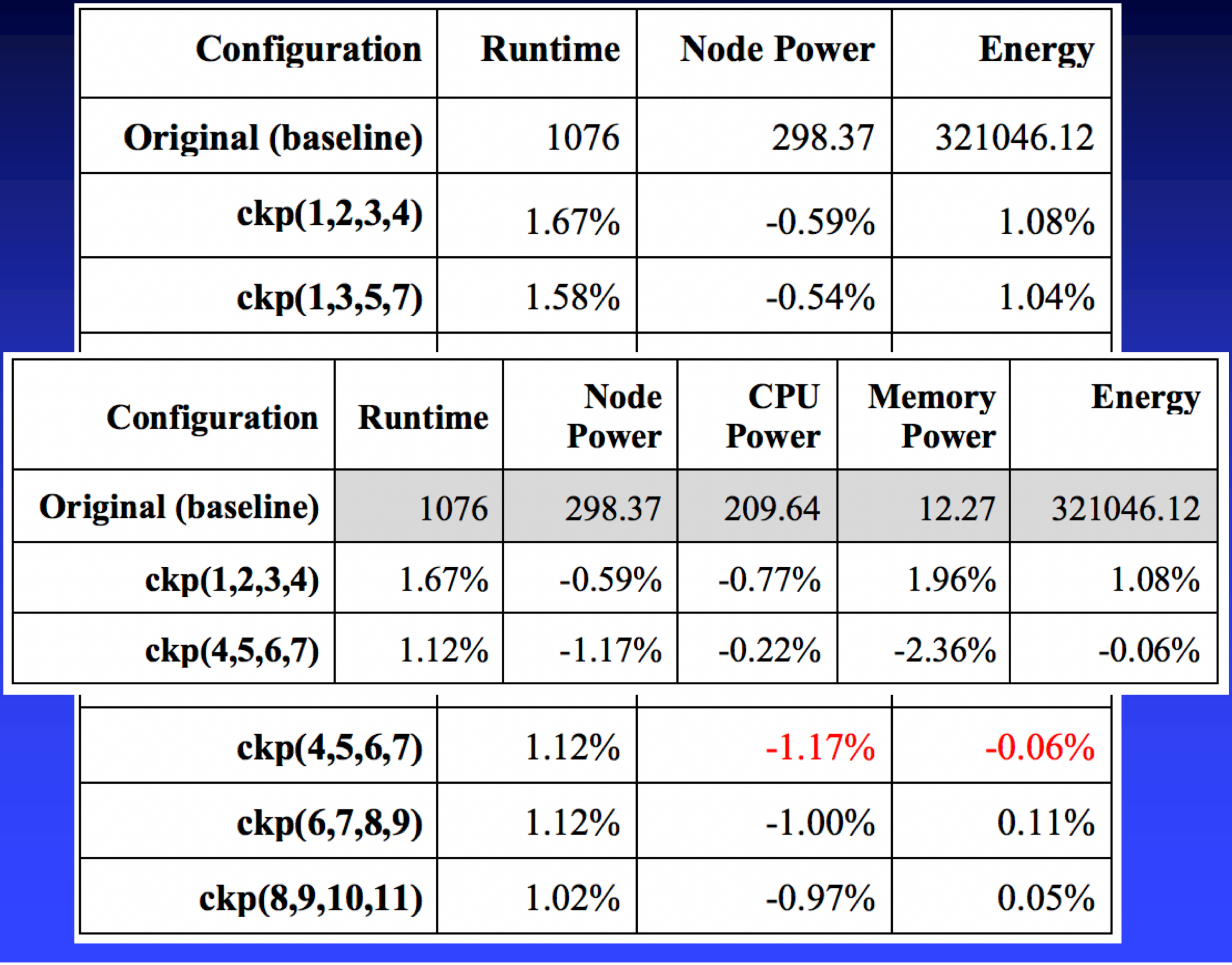
- HDC computes the 2D heat distribution over time based on a set of initial heat sources, and it is compute-intensive
- The checkpointing file size is 32 MB per MPI process

- An FTI application can perform checkpoints with various frequencies and bit-flip failure injections at different bit positions.
  - 10 checkpointing configurations
  - 7 configurations with one bit-flip failure injection
  - 5 different bit positions

Cray XC40 Theta at ANL

System Name	Cray XC40 Theta
Architecture	Intel KNL
Number of nodes	3,624
CPU cores per node	64
Sockets per node	1
CPU type and speed	Xeon Phi KNL 7230 1.30GHz
L1 cache per core	D:32KB/I:32KB
L2 cache per socket	32MB (shared)
L3 cache per socket	None
Memory per node	16GB/192GB
Network	Cray Aries Dragonfly
Power tools	CapMC/PoLiMEr
TDP per socket	215W
Power Management	Yes
File System	Lustre PFS

Results for HDC on Cray XC40

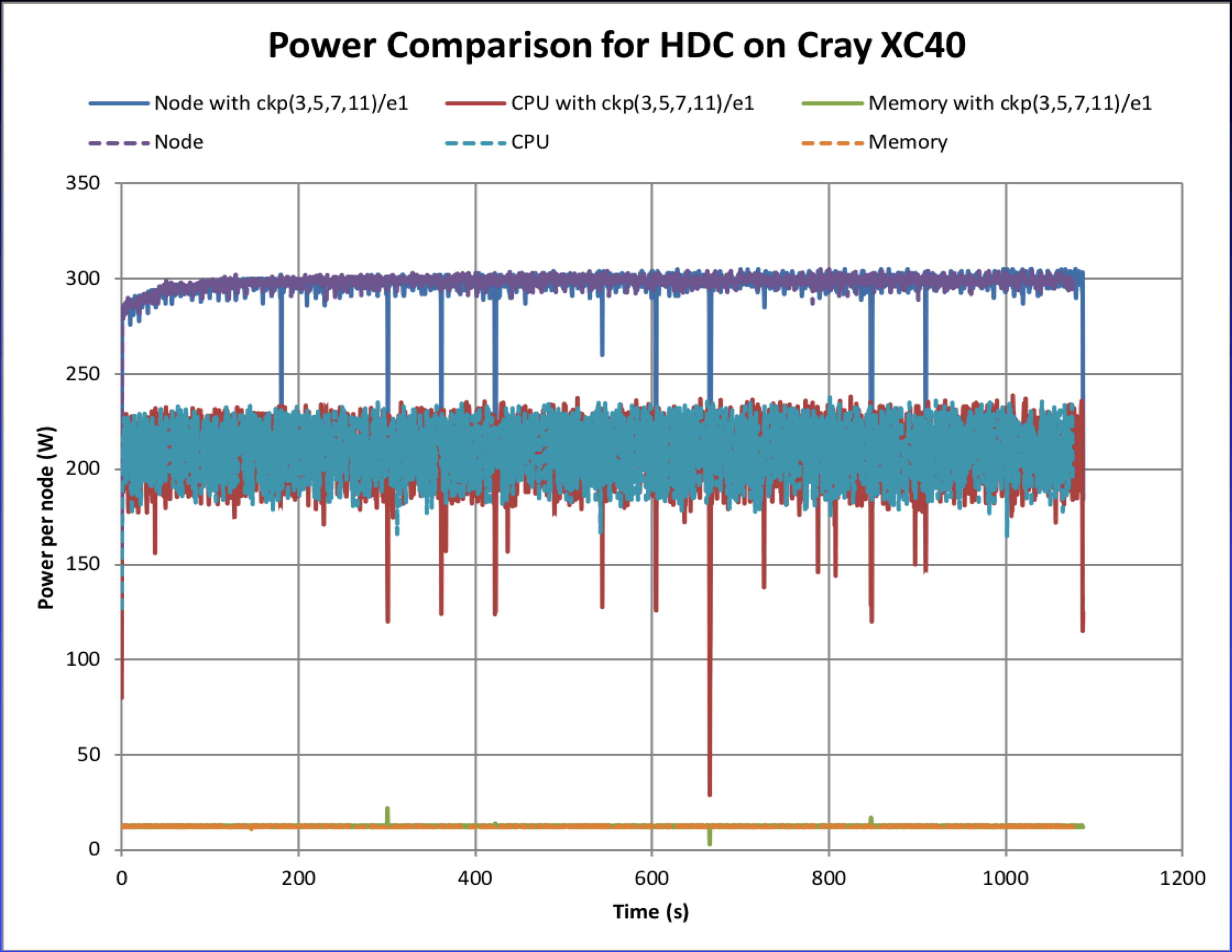


Configuration	Runtime	Node Power	Energy
Original (baseline)	1076	298.37	321046.12
ckp(1,2,3,4)	1.67%	-0.59%	1.08%
ckp(1,3,5,7)	1.58%	-0.54%	1.04%
ckp(4,5,6,7)	1.12%	-1.17%	-0.06%
ckp(6,7,8,9)	1.12%	-1.00%	0.11%
ckp(8,9,10,11)	1.02%	-0.97%	0.05%

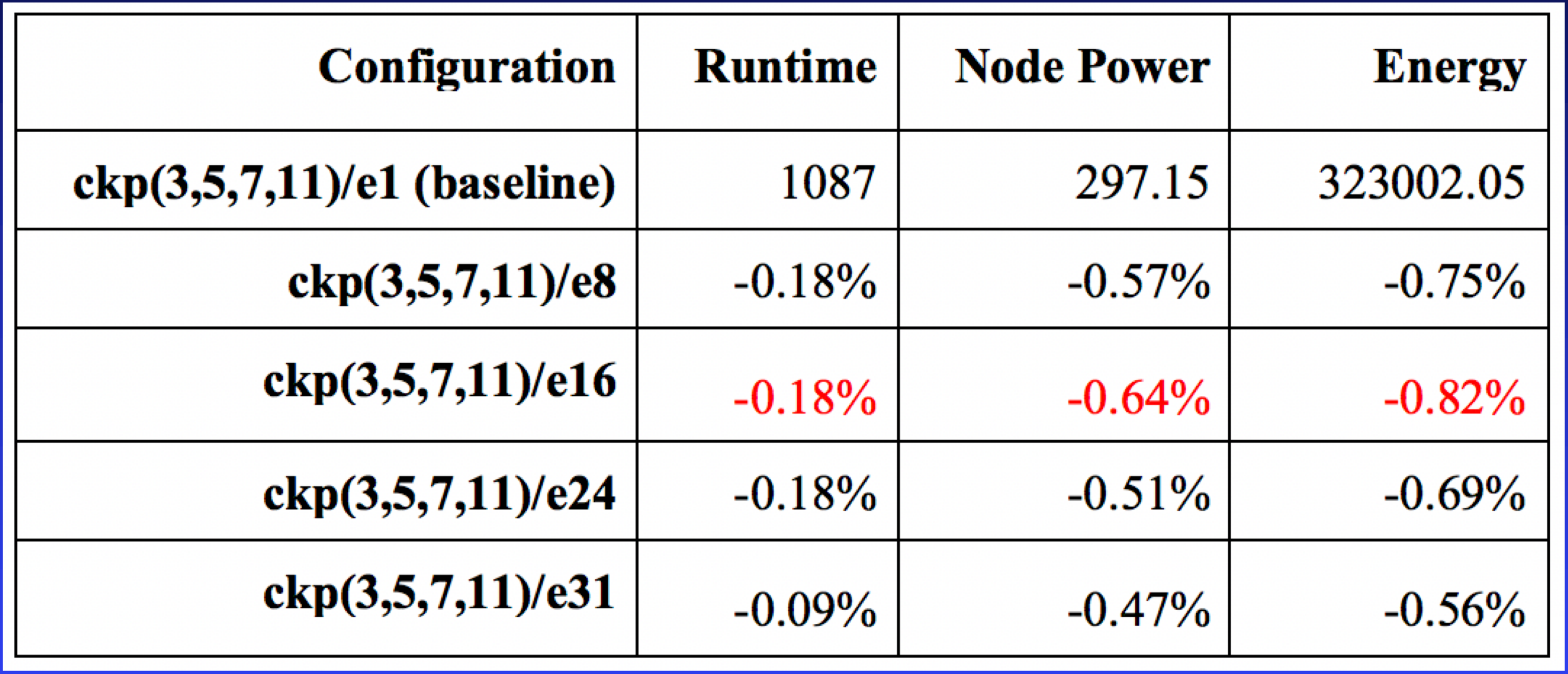
Configuration	Runtime	Node Power	CPU Power	Memory Power	Energy
Original (baseline)	1076	298.37	209.64	12.27	321046.12
ckp(1,2,3,4)	1.67%	-0.59%	-0.77%	1.96%	1.08%
ckp(4,5,6,7)	1.12%	-1.17%	-0.22%	-2.36%	-0.06%

Results for HDC on Cray XC40

Configuration	Runtime	Node Power	Energy
ckp(3,5,7,11)/e1 (baseline)	1087	297.15	323002.05
ckp(1,2,3,4)/e1	0.74%	-0.31%	0.42%
ckp(2,3,4,5)/e1	0.46%	-0.12%	0.33%
ckp(2,4,6,8)/e1	0	-0.58%	-0.58%
ckp(4,5,6,7)/e1	-0.18%	-0.19%	-0.37%
ckp(6,7,8,9)/e1	-0.37%	-0.18%	-0.55%
ckp(8,9,10,11)/e1	-0.55%	-0.04%	-0.60%


Power Comparison for HDC on Cray XC40
Node with ckp(3,5,7,11)/e1
CPU with ckp(3,5,7,11)/e1
Memory with ckp(3,5,7,11)/e1
Node
CPU
Memory
Power per node (W)
350
300
250
200
150
100
50
0
0
200
400
600
800
1000
1200
Time (s)

Results for HDC on Cray XC40



Configuration	Runtime	Node Power	Energy
ckp(3,5,7,11)/e1 (baseline)	1087	297.15	323002.05
ckp(3,5,7,11)/e8	-0.18%	-0.57%	-0.75%
ckp(3,5,7,11)/e16	-0.18%	-0.64%	-0.82%
ckp(3,5,7,11)/e24	-0.18%	-0.51%	-0.69%
ckp(3,5,7,11)/e31	-0.09%	-0.47%	-0.56%

IBM BlueGene/Q Mira at ANL

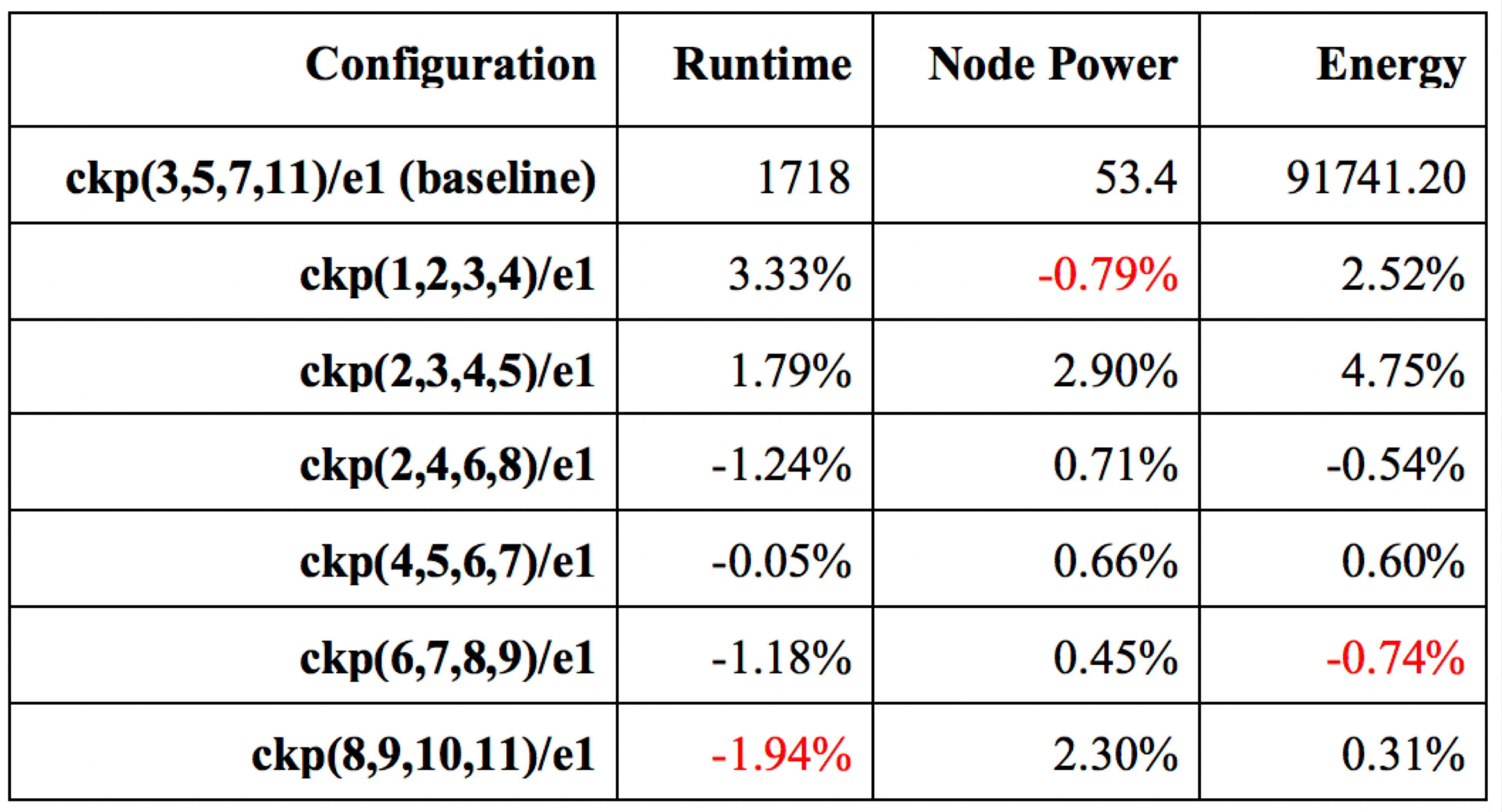
System Name	IBM BG/Q Mira
Architecture	IBM BG/Q
Number of nodes	49,152
CPU cores per node	16
Sockets per node	1
CPU type and speed	PowerPC A2 1.6GHz
L1 cache per core	D:16KB/I:16KB
L2 cache per socket	32MB (shared)
L3 cache per socket	None
Memory per node	16GB
Network	5D Torus
Power tools	EMON/MonEQ
TDP per socket	55W
Power Management	No
File System	GPFS

Results for HDC on IBM BG/Q

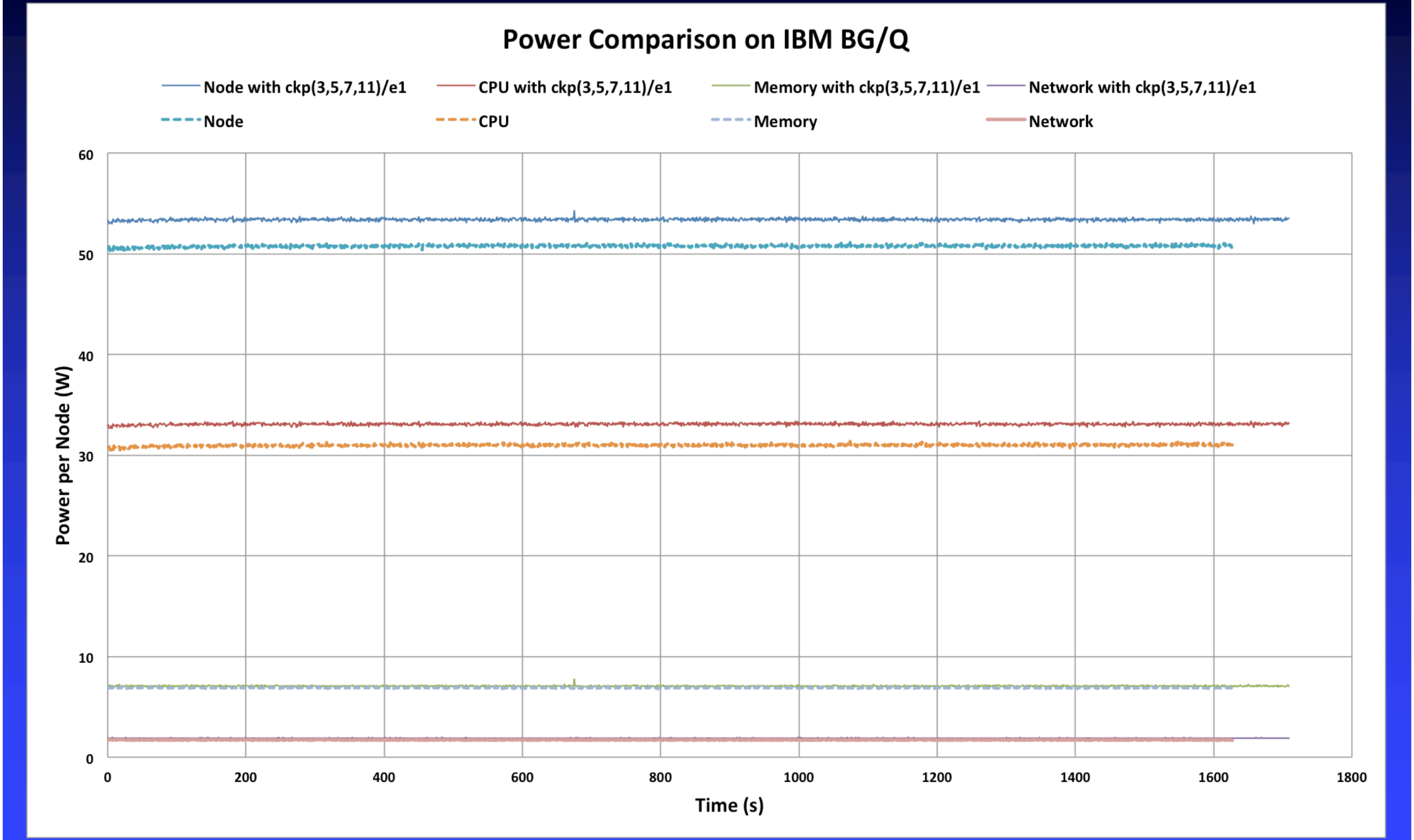
Configuration	Runtime	Node Power	Energy
Original (baseline)	1626	50.77	82552.02
ckp(1,2,3,4)	9.29%	7.52%	17.51%
ckp(1,3,5,7)	8.74%	2.56%	11.53%
ckp(2,3,4,5)	7.44%	3.13%	10.80%
ckp(4,5,6,7)	5.65%	3.22%	9.17%
ckp(6,7,8,9)	8.99%	2.21%	11.39%
ckp(8,9,10,11)	3.32%	3.39%	6.82%

Configuration	Runtime	Node Power	CPU Power	Memory Power	Network Power	Energy
Original (baseline)	1626	50.77	30.99	6.89	1.72	82552.02
ckp(1,2,3,4)	9.29%	7.52%	8.55%	9.14%	8.14%	17.51%
ckp(8,9,10,11)	3.32%	3.39%	1.58%	13.79%	8.14%	6.82%

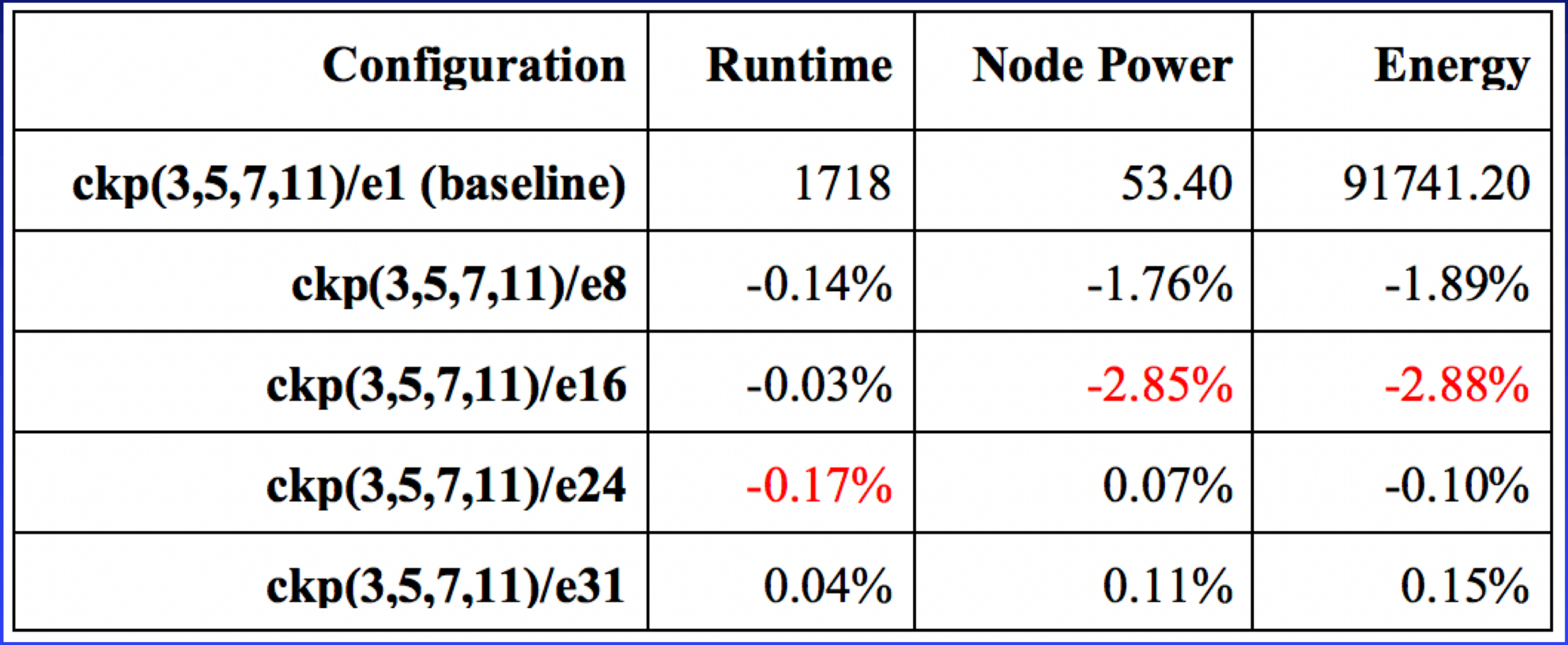
Results for HDC on IBM BG/Q



Configuration	Runtime	Node Power	Energy
ckp(3,5,7,11)/e1 (baseline)	1718	53.4	91741.20
ckp(1,2,3,4)/e1	3.33%	-0.79%	2.52%
ckp(2,3,4,5)/e1	1.79%	2.90%	4.75%
ckp(2,4,6,8)/e1	-1.24%	0.71%	-0.54%
ckp(4,5,6,7)/e1	-0.05%	0.66%	0.60%
ckp(6,7,8,9)/e1	-1.18%	0.45%	-0.74%
ckp(8,9,10,11)/e1	-1.94%	2.30%	0.31%

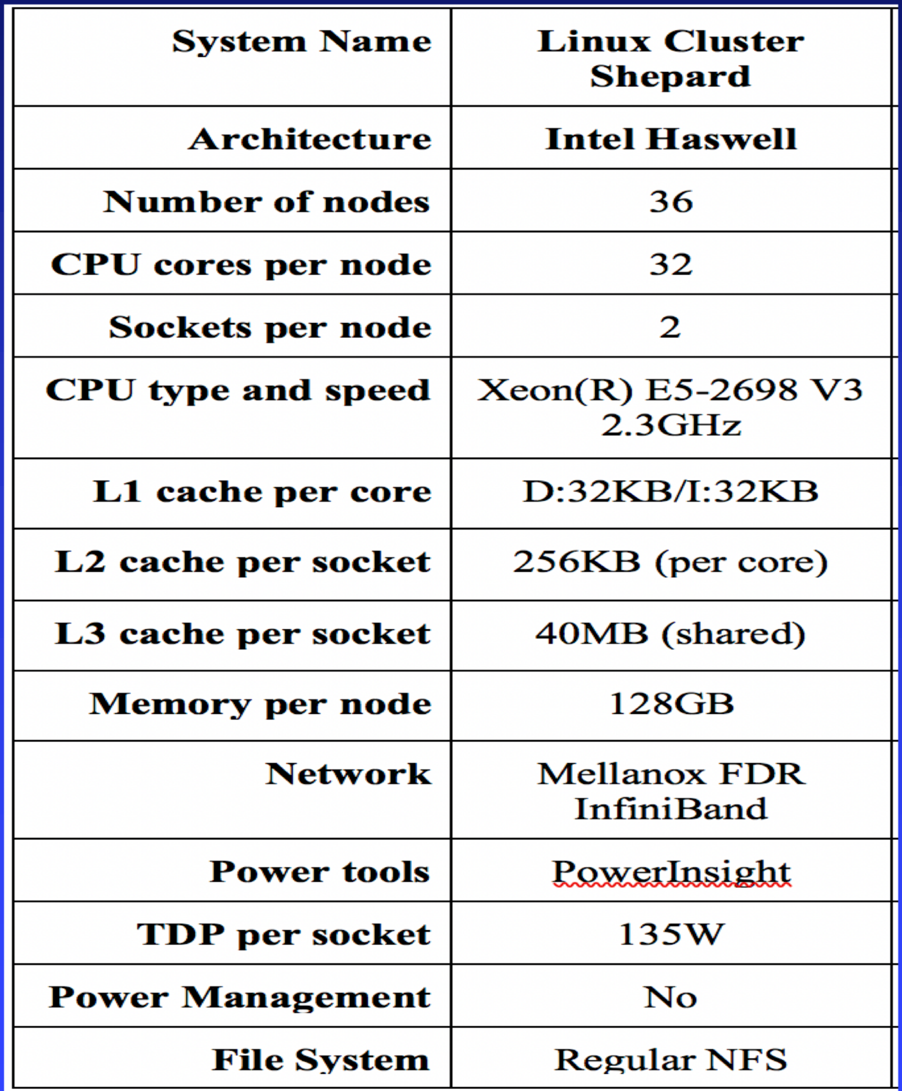

Power Comparison on IBM BG/Q
Node with ckp(3,5,7,11)/e1
CPU with ckp(3,5,7,11)/e1
Memory with ckp(3,5,7,11)/e1
Network with ckp(3,5,7,11)/e1
Node
CPU
Memory
Network
Power per Node (W)
60
50
40
30
20
10
0
0
200
400
600
800
1000
1200
1400
1600
1800
Time (s)

Results for HDC on IBM BG/Q



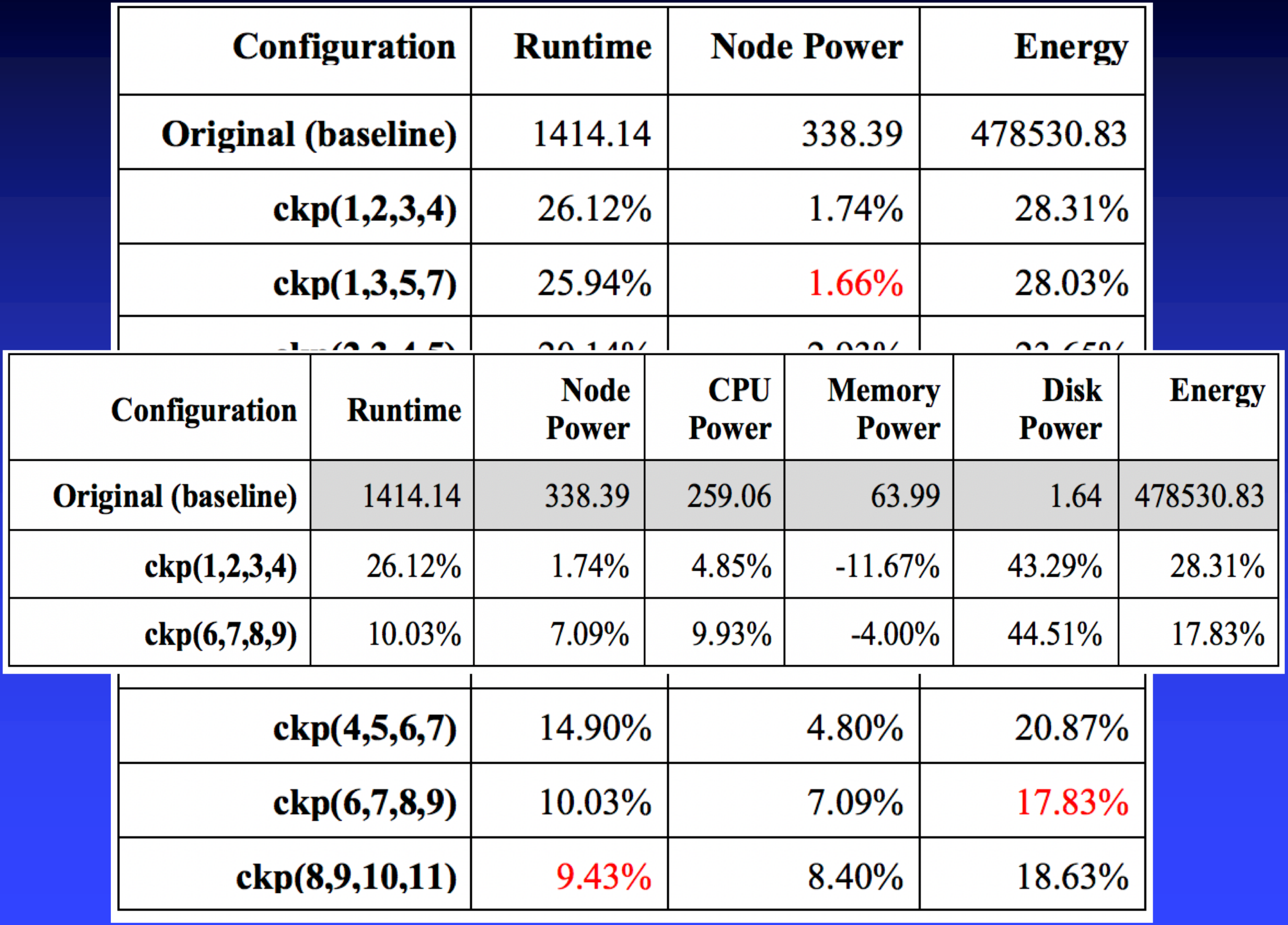
Configuration	Runtime	Node Power	Energy
ckp(3,5,7,11)/e1 (baseline)	1718	53.40	91741.20
ckp(3,5,7,11)/e8	-0.14%	-1.76%	-1.89%
ckp(3,5,7,11)/e16	-0.03%	-2.85%	-2.88%
ckp(3,5,7,11)/e24	-0.17%	0.07%	-0.10%
ckp(3,5,7,11)/e31	0.04%	0.11%	0.15%

Intel Haswell cluster Shepard at SNL



System Name	Linux Cluster Shepard
Architecture	**Intel Haswell**
Number of nodes	36
CPU cores per node	32
Sockets per node	2
CPU type and speed	Xeon(R) E5-2698 V3 2.3GHz
L1 cache per core	D:32KB/I:32KB
L2 cache per socket	256KB (per core)
L3 cache per socket	40MB (shared)
Memory per node	128GB
Network	Mellanox FDR InfiniBand
Power tools	PowerInsight
TDP per socket	135W
Power Management	No
File System	Regular NFS

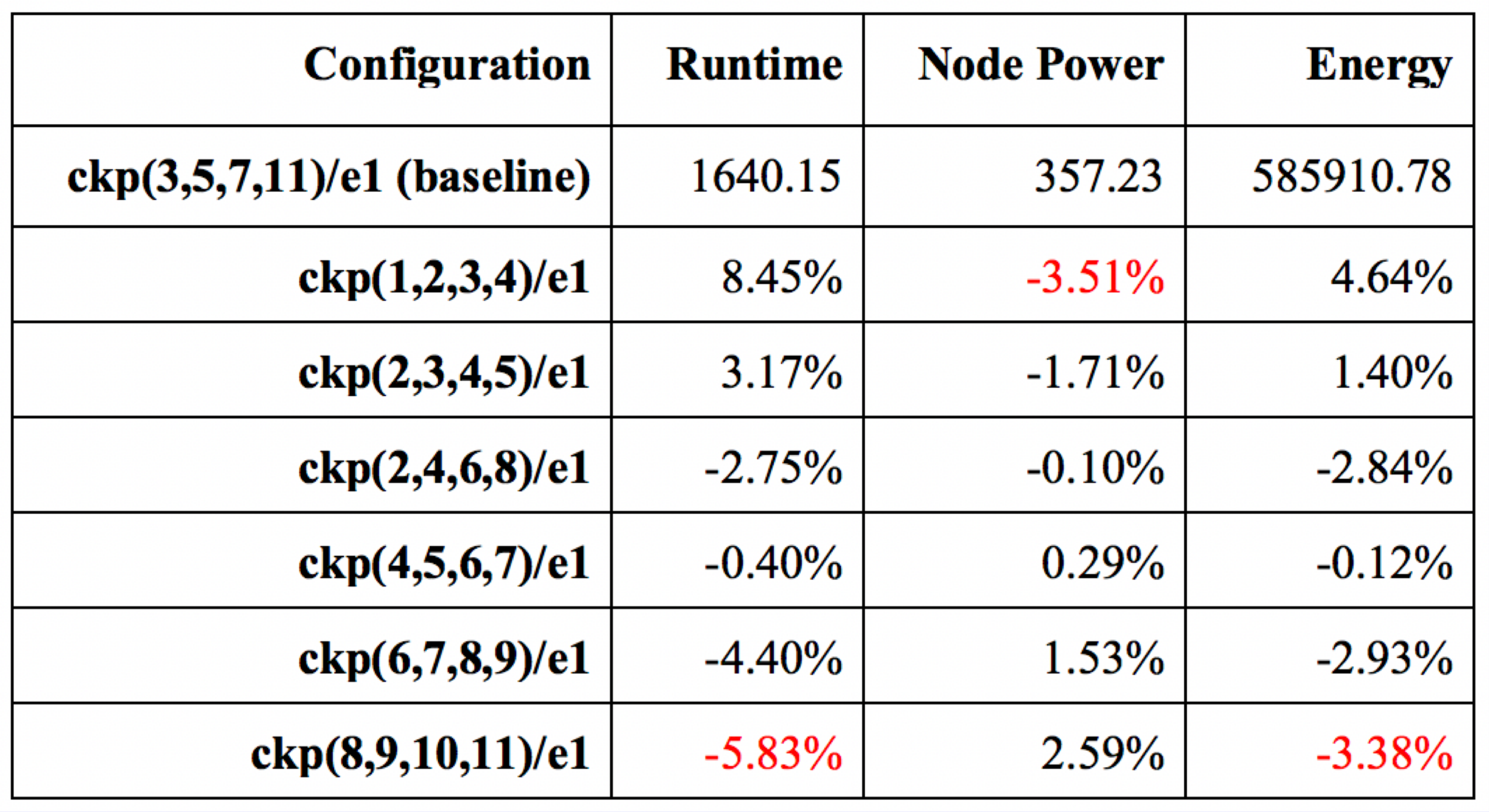
Results for HDC on Intel Haswell



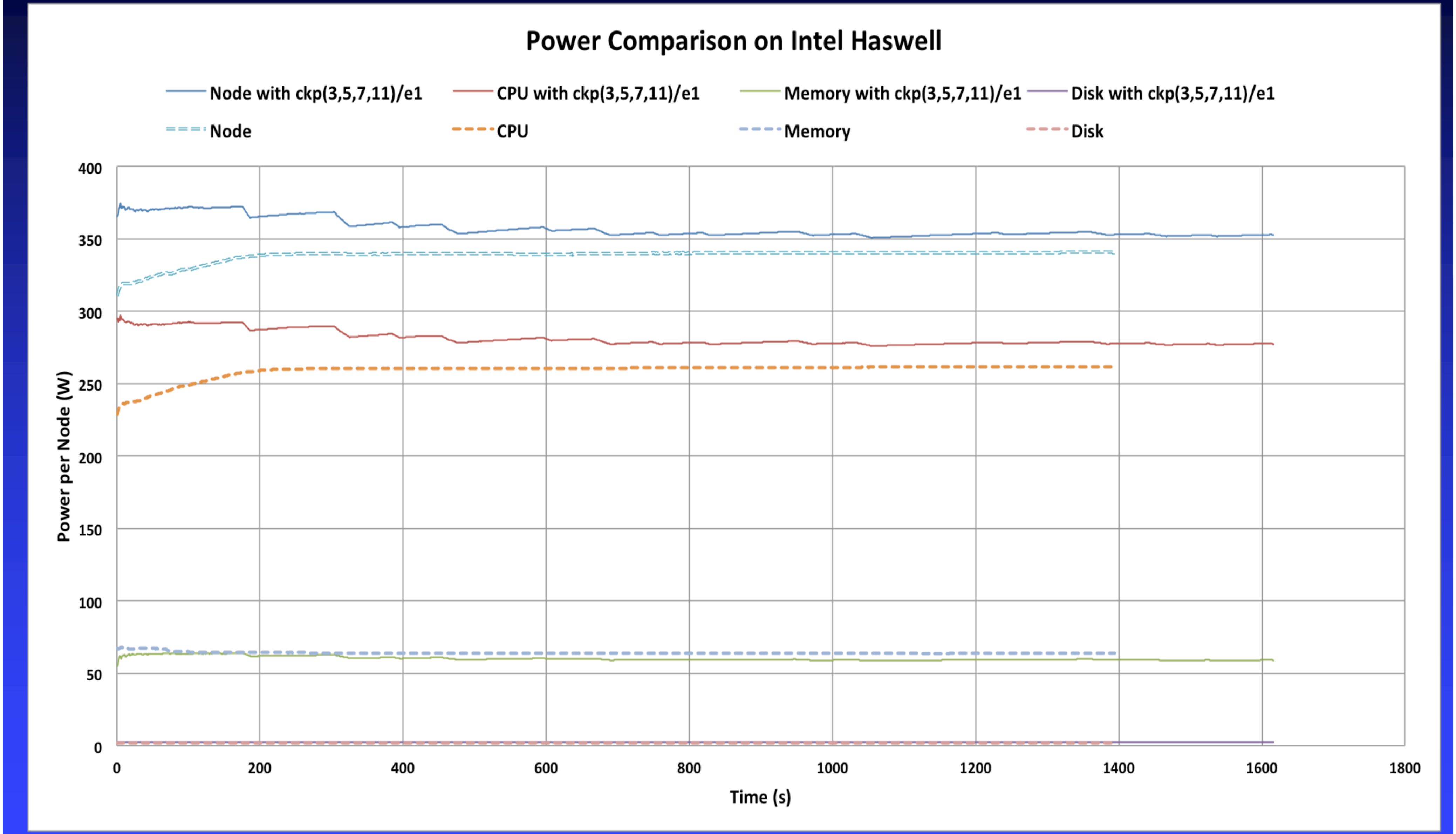
Configuration	Runtime	Node Power	Energy
Original (baseline)	1414.14	338.39	478530.83
ckp(1,2,3,4)	26.12%	1.74%	28.31%
ckp(1,3,5,7)	25.94%	1.66%	28.03%
ckp(2,3,4,5)	20.14%	2.93%	23.65%
ckp(4,5,6,7)	14.90%	4.80%	20.87%
ckp(6,7,8,9)	10.03%	7.09%	17.83%
ckp(8,9,10,11)	9.43%	8.40%	18.63%

Configuration	Runtime	Node Power	CPU Power	Memory Power	Disk Power	Energy
Original (baseline)	1414.14	338.39	259.06	63.99	1.64	478530.83
ckp(1,2,3,4)	26.12%	1.74%	4.85%	-11.67%	43.29%	28.31%
ckp(6,7,8,9)	10.03%	7.09%	9.93%	-4.00%	44.51%	17.83%

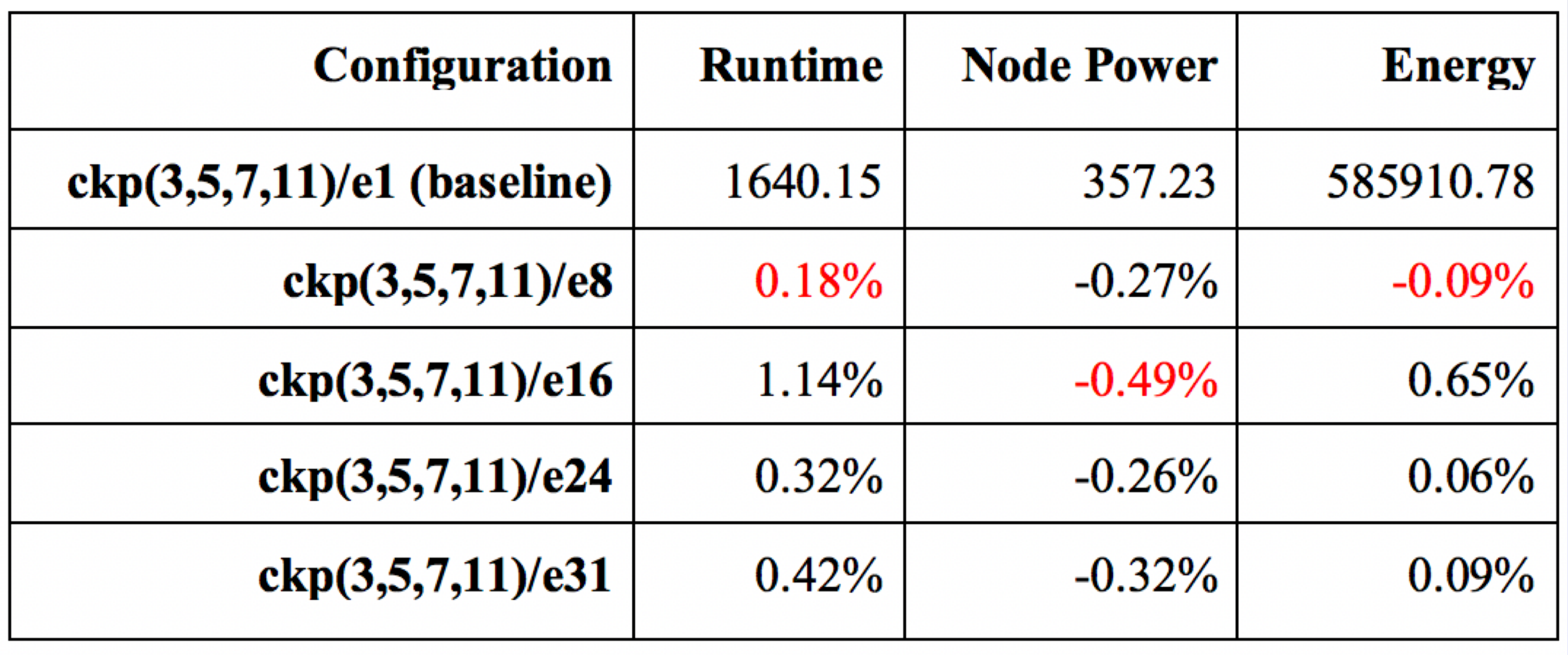
Results for HDC on Intel Haswell



Configuration	Runtime	Node Power	Energy
ckp(3,5,7,11)/e1 (baseline)	1640.15	357.23	585910.78
ckp(1,2,3,4)/e1	8.45%	-3.51%	4.64%
ckp(2,3,4,5)/e1	3.17%	-1.71%	1.40%
ckp(2,4,6,8)/e1	-2.75%	-0.10%	-2.84%
ckp(4,5,6,7)/e1	-0.40%	0.29%	-0.12%
ckp(6,7,8,9)/e1	-4.40%	1.53%	-2.93%
ckp(8,9,10,11)/e1	-5.83%	2.59%	-3.38%

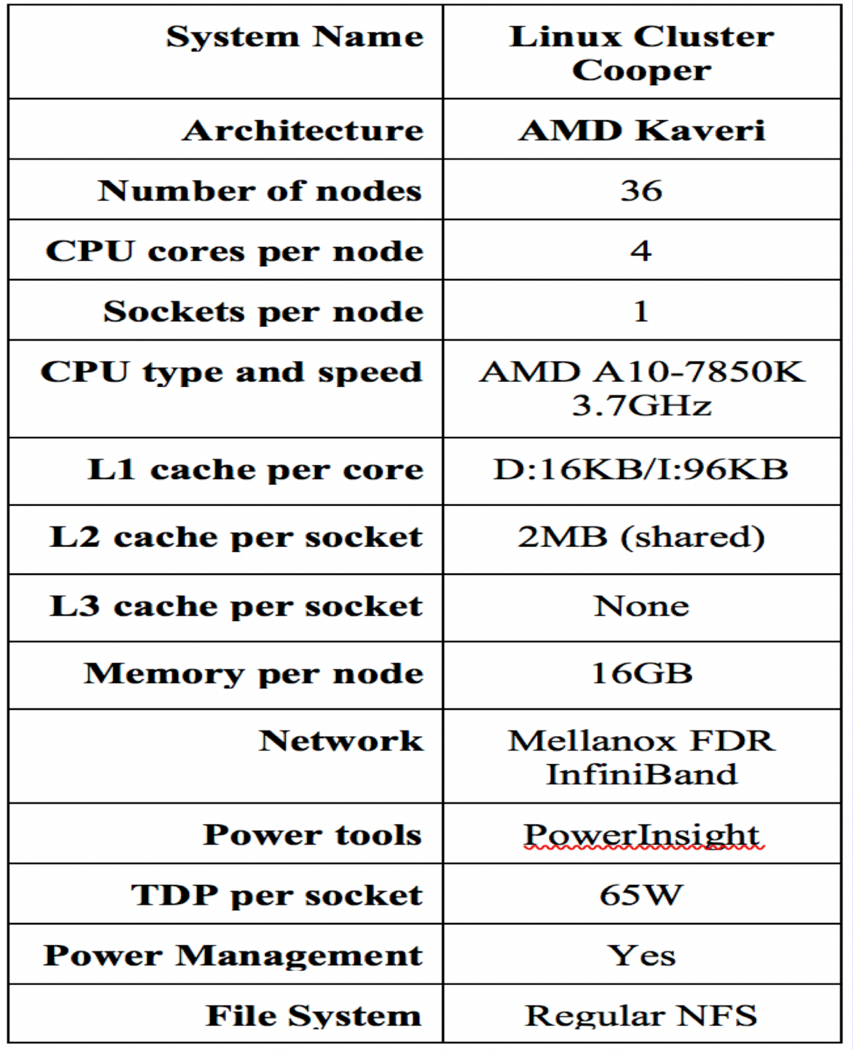

Power Comparison on Intel Haswell
Node with ckp(3,5,7,11)/e1
CPU with ckp(3,5,7,11)/e1
Memory with ckp(3,5,7,11)/e1
Disk with ckp(3,5,7,11)/e1
Node
CPU
Memory
Disk
Power per Node (W)
400
350
300
250
200
150
100
50
0
0
200
400
600
800
1000
1200
1400
1600
1800
Time (s)

Results for HDC on Intel Haswell



Configuration	Runtime	Node Power	Energy
ckp(3,5,7,11)/e1 (baseline)	1640.15	357.23	585910.78
ckp(3,5,7,11)/e8	0.18%	-0.27%	-0.09%
ckp(3,5,7,11)/e16	1.14%	-0.49%	0.65%
ckp(3,5,7,11)/e24	0.32%	-0.26%	0.06%
ckp(3,5,7,11)/e31	0.42%	-0.32%	0.09%

AMD Kaveri Cluster Cooper at SNL



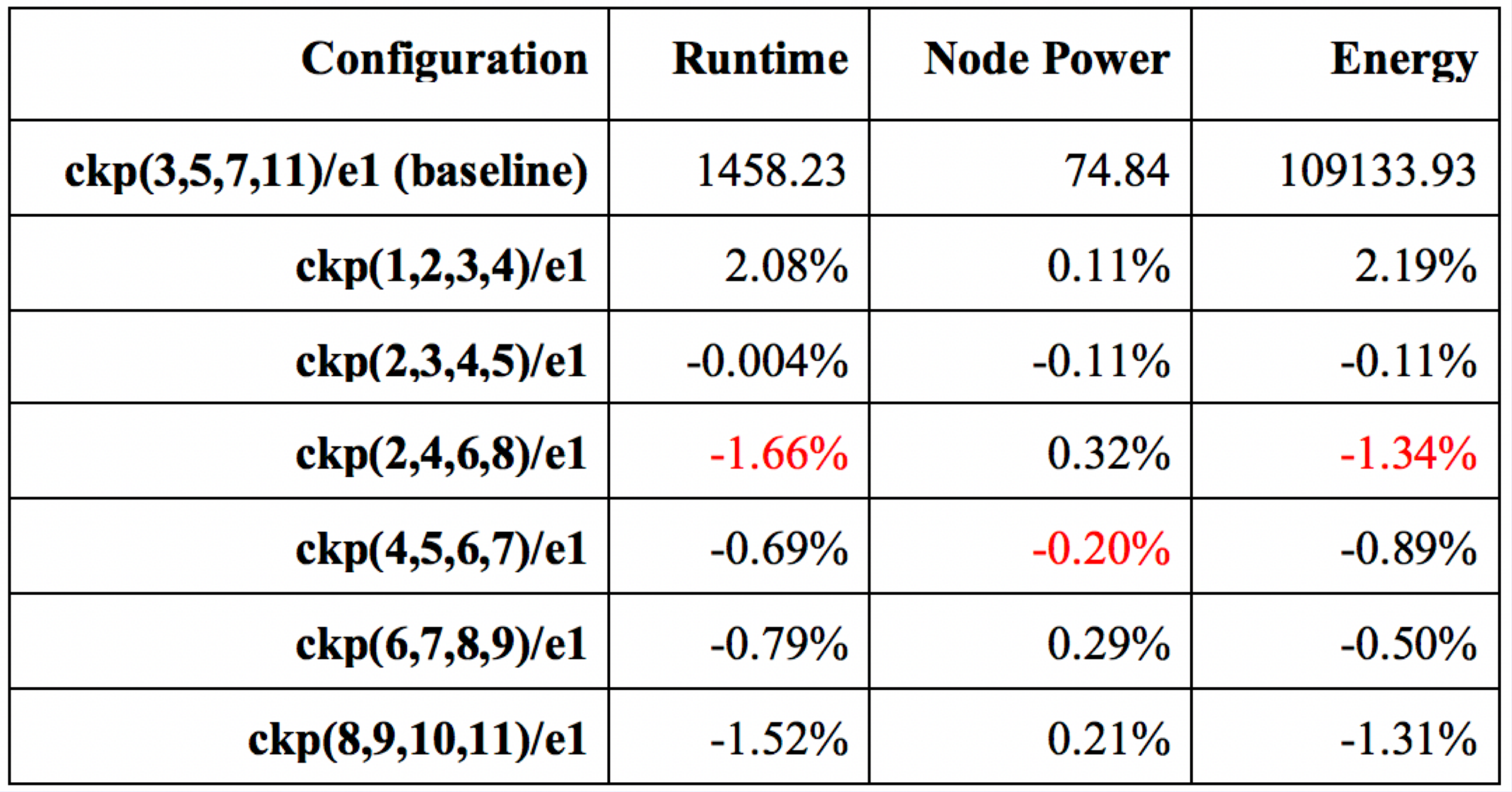
System Name	**Linux Cluster Cooper**
Architecture	**AMD Kaveri**
Number of nodes	36
CPU cores per node	4
Sockets per node	1
CPU type and speed	AMD A10-7850K 3.7GHz
L1 cache per core	D:16KB/I:96KB
L2 cache per socket	2MB (shared)
L3 cache per socket	None
Memory per node	16GB
Network	Mellanox FDR InfiniBand
Power tools	PowerInsight
TDP per socket	65W
Power Management	Yes
File System	Regular NFS

Results for HDC on AMD Kaveri

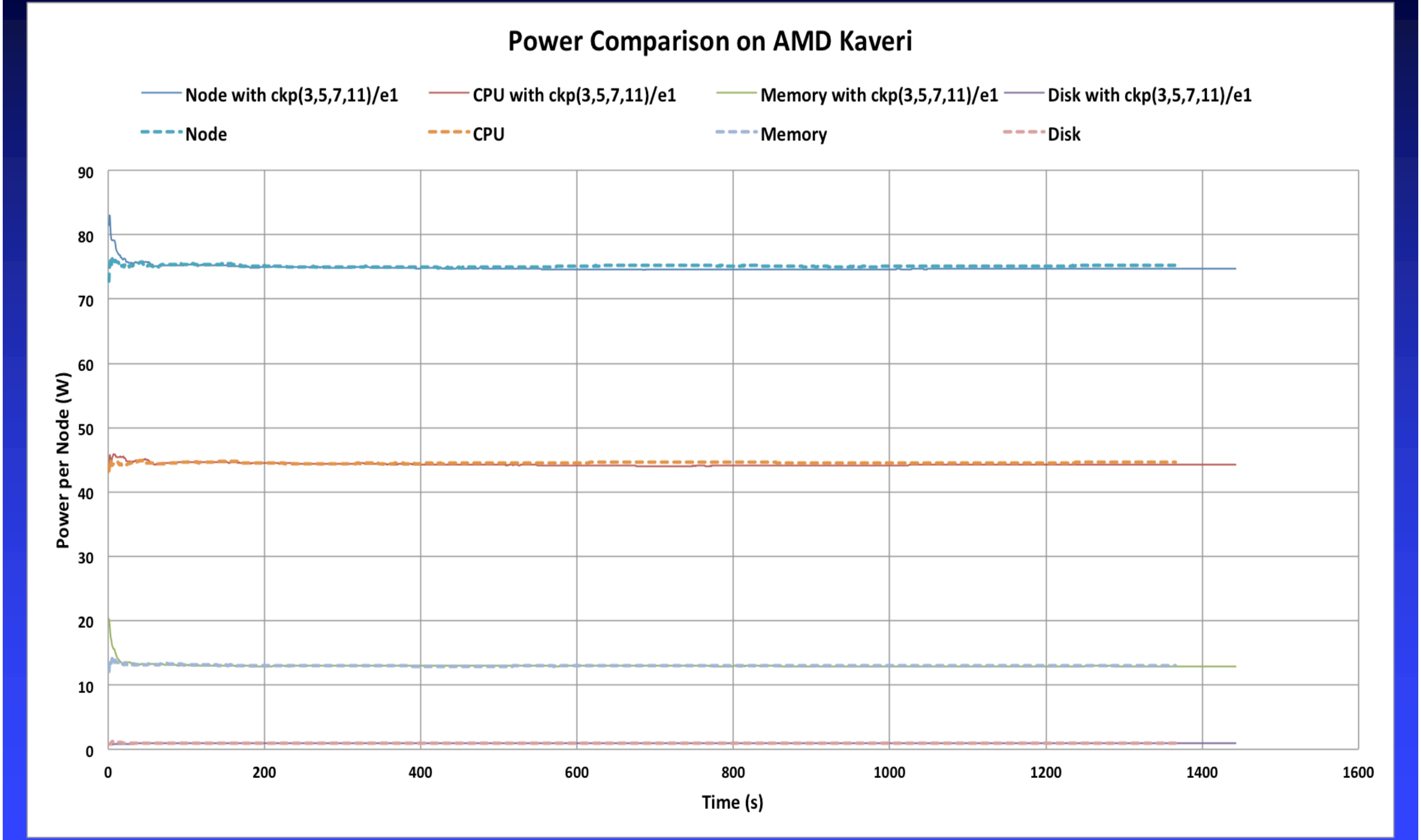
Configuration	Runtime	Node Power	Energy
Original (baseline)	1380.92	75.14	103762.33
ckp(1,2,3,4)	7.15%	-0.37%	6.75%
ckp(1,3,5,7)	6.30%	-0.07%	6.23%
ckp(2,3,4,5)	5.64%	0.15%	5.79%
ckp(4,5,6,7)	5.43%	0.32%	5.77%
ckp(6,7,8,9)	3.93%	-0.27%	3.65%
ckp(8,9,10,11)	3.15%	0.45%	3.62%

Configuration	Runtime	Node Power	CPU Power	Memory Power	Disk Power	Energy
Original (baseline)	1380.92	75.14	44.54	13.01	0.92	103762.33
ckp(1,2,3,4)	7.15%	-0.37%	-0.47%	-0.38%	1.09%	6.75%
ckp(8,9,10,11)	3.15%	0.45%	0.36%	0.0%	0.0%	3.62%

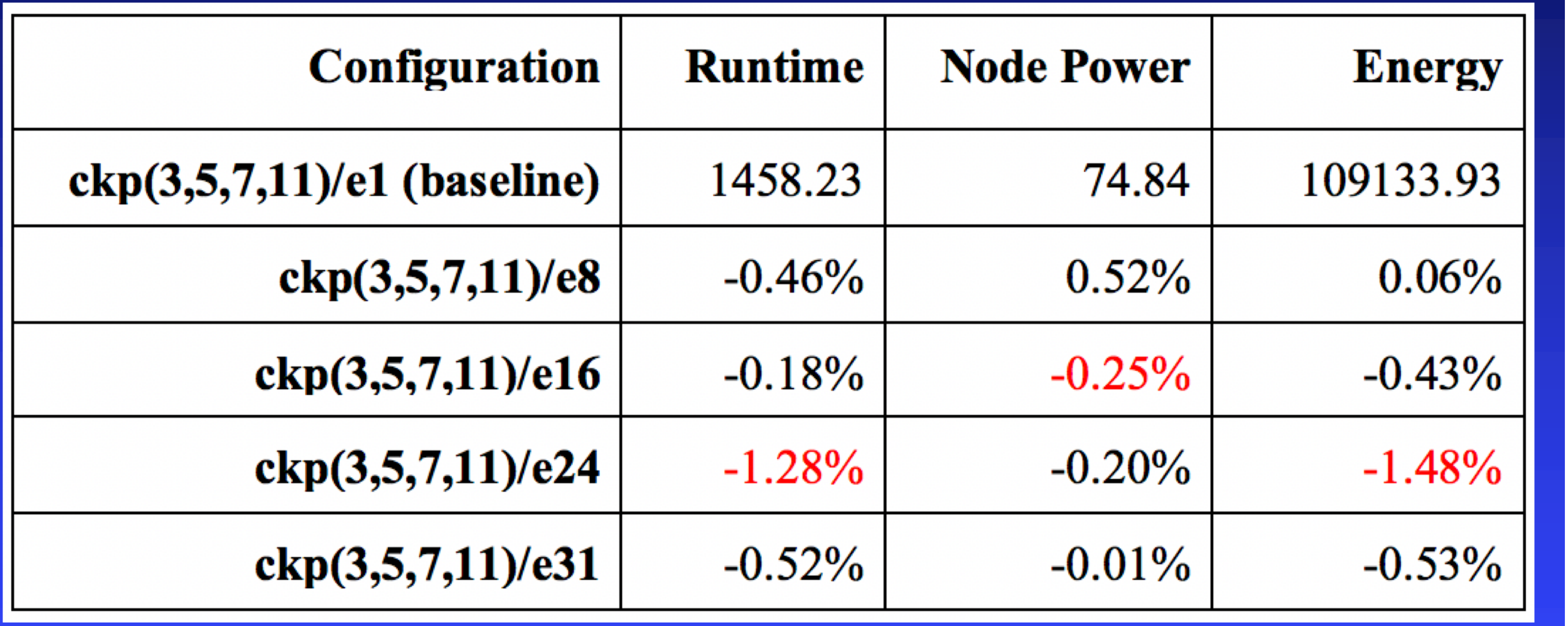
Results for HDC on AMD Kaveri



Configuration	Runtime	Node Power	Energy
ckp(3,5,7,11)/e1 (baseline)	1458.23	74.84	109133.93
ckp(1,2,3,4)/e1	2.08%	0.11%	2.19%
ckp(2,3,4,5)/e1	-0.004%	-0.11%	-0.11%
ckp(2,4,6,8)/e1	-1.66%	0.32%	-1.34%
ckp(4,5,6,7)/e1	-0.69%	-0.20%	-0.89%
ckp(6,7,8,9)/e1	-0.79%	0.29%	-0.50%
ckp(8,9,10,11)/e1	-1.52%	0.21%	-1.31%

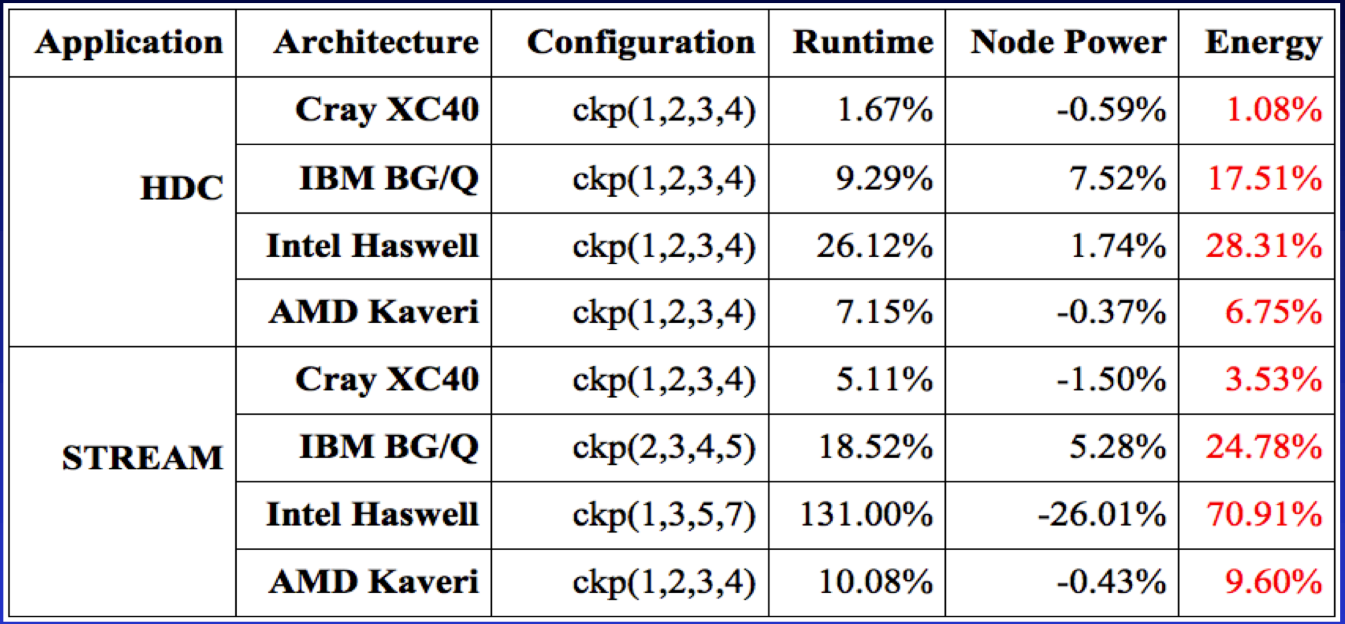

Power Comparison on AMD Kaveri
Node with ckp(3,5,7,11)/e1
CPU with ckp(3,5,7,11)/e1
Memory with ckp(3,5,7,11)/e1
Disk with ckp(3,5,7,11)/e1
Node
CPU
Memory
Disk
Power per Node (W)
90
80
70
60
50
40
30
20
10
0
0
200
400
600
800
1000
1200
1400
1600
Time (s)

Results for HDC on AMD Kaveri



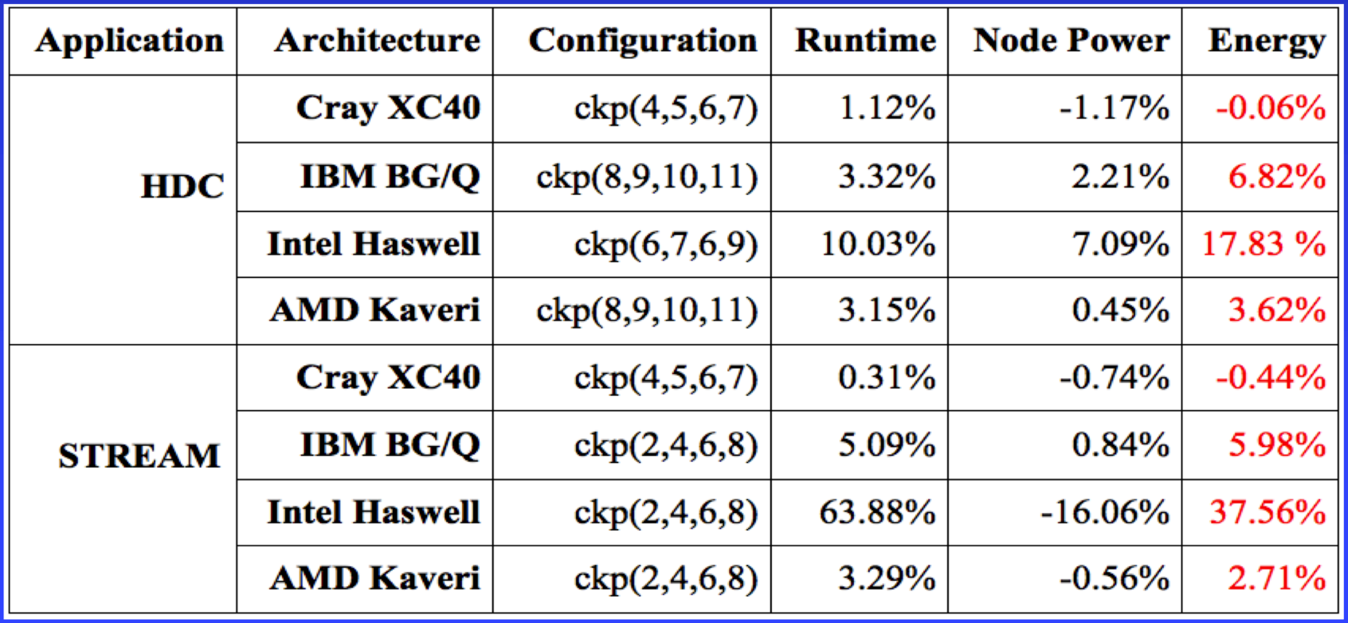
Configuration	Runtime	Node Power	Energy
ckp(3,5,7,11)/e1 (baseline)	1458.23	74.84	109133.93
ckp(3,5,7,11)/e8	-0.46%	0.52%	0.06%
ckp(3,5,7,11)/e16	-0.18%	-0.25%	-0.43%
ckp(3,5,7,11)/e24	-1.28%	-0.20%	-1.48%
ckp(3,5,7,11)/e31	-0.52%	-0.01%	-0.53%

Summary of Results for HDC and STREAM



Application	Architecture	Configuration	Runtime	Node Power	Energy
HDC	Cray XC40	ckp(1,2,3,4)	1.67%	-0.59%	1.08%
	IBM BG/Q	ckp(1,2,3,4)	9.29%	7.52%	17.51%
	Intel Haswell	ckp(1,2,3,4)	26.12%	1.74%	28.31%
	AMD Kaveri	ckp(1,2,3,4)	7.15%	-0.37%	6.75%
STREAM	Cray XC40	ckp(1,2,3,4)	5.11%	-1.50%	3.53%
	IBM BG/Q	ckp(2,3,4,5)	18.52%	5.28%	24.78%
	Intel Haswell	ckp(1,3,5,7)	131.00%	-26.01%	70.91%
	AMD Kaveri	ckp(1,2,3,4)	10.08%	-0.43%	9.60%

Maximum Energy



Application	Architecture	Configuration	Runtime	Node Power	Energy
HDC	Cray XC40	ckp(4,5,6,7)	1.12%	-1.17%	-0.06%
	IBM BG/Q	ckp(8,9,10,11)	3.32%	2.21%	6.82%
	Intel Haswell	ckp(6,7,6,9)	10.03%	7.09%	17.83 %
	AMD Kaveri	ckp(8,9,10,11)	3.15%	0.45%	3.62%
STREAM	Cray XC40	ckp(4,5,6,7)	0.31%	-0.74%	-0.44%
	IBM BG/Q	ckp(2,4,6,8)	5.09%	0.84%	5.98%
	Intel Haswell	ckp(2,4,6,8)	63.88%	-16.06%	37.56%
	AMD Kaveri	ckp(2,4,6,8)	3.29%	-0.56%	2.71%

Minimum Energy

Summary

- The difference between maximum and minimum energy percentages is 4% on Cray XC40, 19% on IBM BG/Q, 34% on Intel Haswell, and 7% on AMD Kaveri

- Both Cray XC40 and AMD Kaveri with dynamic power management exhibited the smallest impact, whereas Intel Haswell without dynamic power management manifested the largest impact

- Bit-flip fault injection had little impact on application runtime and power consumption

Future Work

- Using DVFS/Power Capping to improve power consumption for FTI-based MPI applications
  - Large FTI overhead during checkpointing for large scale scientific applications
  - Apply DVFS/Power Capping to the stage

- Investigating other fault tolerance protocols
  - ABFT (Algorithm-based Fault Tolerance)
    - FTLA (QR, LU) developed by UTK