Towards an Energy Estimator for Fault Tolerance Protocols

Mohammed el Mehdi Diouri[1], Olivier Glück[1], Laurent Lefèvre[1] and Franck Cappello[2]
1. **INRIA Avalon Team, ENS Lyon, and Université de Lyon, France.**
2. **INRIA and University of Illinois at Urbana Champaign**
{mehdi.diouri, olivier.gluck, laurent.lefevre}@ens-lyon.fr, cappello@illinois.edu

I – Context and Motivations

Fault free execution of coordinated and uncoordinated checkpointing.

- Checkpointing: storing a snapshot image of the current application state.
- Coordination: synchronizing the processes before taking the checkpoints.
- Message logging: saving on each sender process the messages sent on a storage media.

Estimate the energy consumption of a particular **fault tolerant protocol** for a large variety of execution configurations.

⇒ Such estimations can be used to **compare FT protocols in terms of energy consumption**.

II – Calibration approach

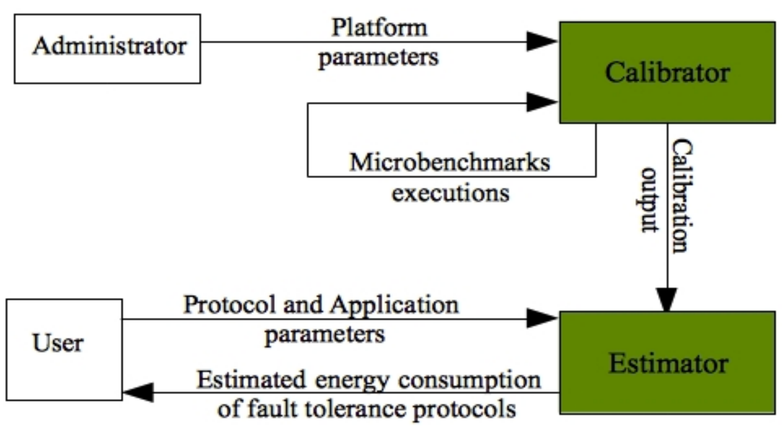
A high level operation: message logging, coordination and checkpointing.
Its energy consumption depends on a large set of parameters **=> Calibration approach.**
Set of simple benchmarks that extracts e^i_{op} for each node *i*

Energy consumption of a high level operation *op* : $e^i_{op} = p^i_{op} \cdot t^i_{op}$

Power consumption of *op* : $p^i_{op} = p^i_{idle} + \Delta p^i_{op}$

Execution time t_{op} :
- Message logging: $t^i_{logging} = t^i_{access} + t^i_{transfer} = t^i_{access} + V_{data} / R^i_{transfer}$
- Coordination: $t^i_{coordination} = t_{synchro} + t^i_{polling} = t_{synchro} + V_{data} / R^i_{transfer}$
- Checkpointing: $t^i_{checkpoint} = t^i_{access} + t^i_{transfer} = t^i_{access} + V_{data} / R^i_{transfer}$

III – Estimation methodology



From the user, protocol and application parameters:
- the total memory size required by the application
- the total number of nodes *N* and the number of processes per node *p*.
- the number of checkpoints *C*
- the total number and size of the messages sent during the application.

From this information, the estimator computes:
- the mean memory size V^{mean}_{mem} required by each node
- the mean volume of data sent V^{mean}_{data} sent (so logged) by each node
- the mean message size $V^{mean}_{message}$

From the calibrator, calibration output:
- $t^i_{checkpoint}$ corresponding to V^{mean}_{mem}
- $t^i_{logging}$ corresponding to V^{mean}_{data}
- $t_{synchro}$ corresponding to *p* and *N* and $t^i_{polling}$ corresponding to $V^{mean}_{message}$

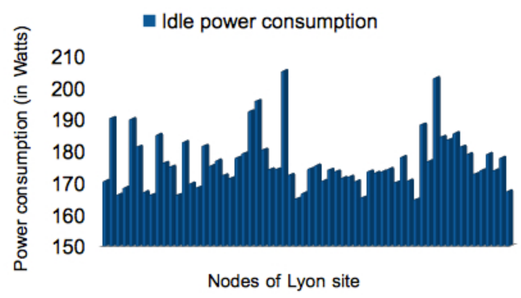
Estimation based on the method of least squares.

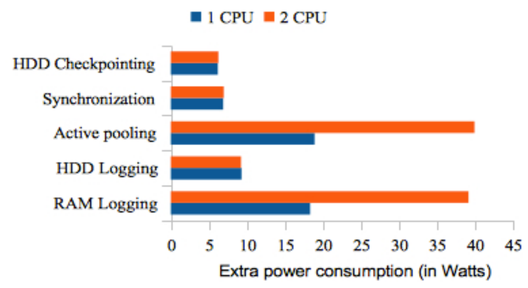
Estimated energy consumption for *op*: $E_{op} = \sum_{i=1}^{N} e^i_{op} = \sum_{i=1}^{N} p^i_{op} \cdot t^i_{op}$

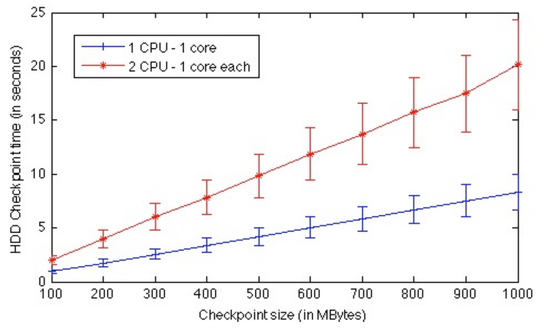
IV – Validation

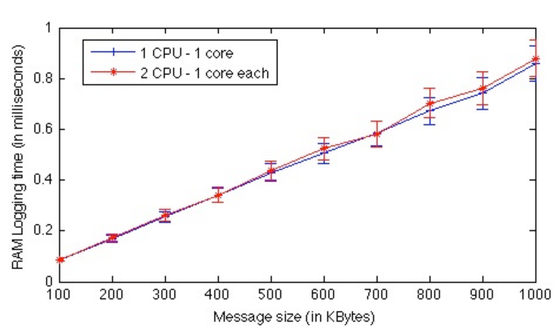
64 available identical nodes Sun Fire V20z where each node gathers:
- 2 AMD Opteron 250 CPU 2.4 GHz, with 1 core each.
- 2 GB of memory; a Gigabit Ethernet network; 73 GB of hard disk drive.

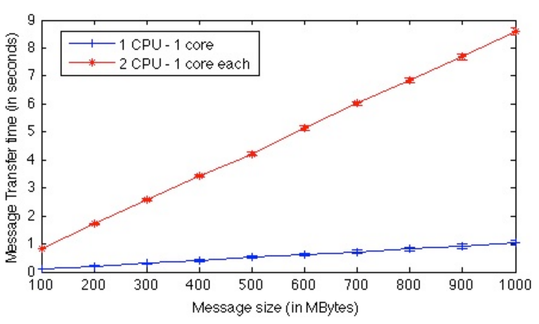
An energy-sensing infrastructure of external power meters from the SME Omegawatt.











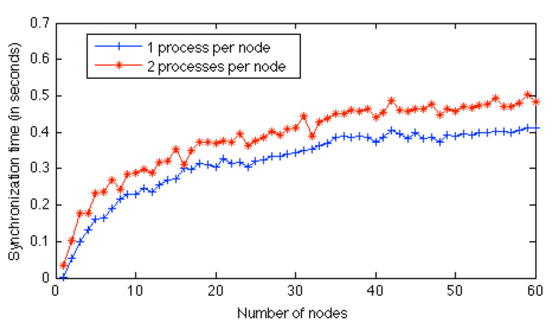


Table 1. Results: Estimated energy (in KJ) — Measured energy (in KJ) — Relative difference in percentage

Application	Context	RAM Logging	HDD Logging	Coordinations	Checkpoints
BT	1 CPU	15.71 — 16.06 — 2.2 %	52.62 — 54.46 — 3.4 %	18.69 — 20.32 — 8.0 %	128.67 — 119.48 — 7.7 %
	2 CPU	9.17 — 8.76 — 4.7 %	52.16 — 50.13 — 4.0 %	10.38 — 11.14 — 6.8 %	158.23 — 152.79 — 3.6 %
CG	1 CPU	14.72 — 14.44 — 1.9 %	45.84 — 47.55 — 3.6 %	14.82 — 15.14 — 2.1 %	49.86 — 51.35 — 2.9 %
	2 CPU	8.20 — 8.67 — 5.4 %	43.36 — 46.10 — 5.9 %	7.98 — 8.15 — 2.0 %	39.45 — 37.75 — 4.5 %
LU	1 CPU	5.38 — 5.85 — 8.0 %	19.31 — 20.09 — 3.9 %	13.95 — 13.18 — 5.8 %	94.29 — 89.34 — 5.5 %
	2 CPU	3.15 — 3.11 — 1.3 %	16.93 — 17.67 — 4.2 %	7.63 — 7.90 — 3.4 %	121.45 — 115.20 — 5.4 %
SP	1 CPU	27.48 — 25.65 — 7.1 %	91.36 — 87.58 — 4.3 %	18.07 — 16.52 — 9.4 %	382.13 — 367.19 — 4.1 %
	2 CPU	17.32 — 18.56 — 6.7 %	82.52 — 86.35 — 4.4 %	11.42 — 10.96 — 4.2 %	246.49 — 263.15 — 6.3 %
CM1	1 CPU	10.18 — 10.91 — 6.7 %	36.78 — 34.21 — 7.5 %	26.86 — 25.12 — 6.9 %	236.41 — 243.12 — 2.8 %
	2 CPU	6.35 — 5.98 — 6.2 %	34.47 — 33.12 — 4.1 %	14.48 — 15.51 — 6.7 %	221.89 — 208.76 — 6.3 %

The relative difference is equal to 4.48 % in average and do not exceed 9.4 %.

IV – Future works

- Investigate energy efficient solutions for FT protocols;
- Include estimations of the recovery
- Validation on a large scale distributed platform (many-cores);
- Estimate more protocols that are needed at extreme-scale:

PPoPP'13, 18th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming.
February 23-27 2013, Shenzhen, China.