SCALABLE DISTRIBUTED COMPUTATION OF CRITICAL PATH

Pierre-Frédérick DENYS
Thursday 8 December 2022

Agenda

- Introduction
- Proposed algorithm
- Benchmarks
- Future work and usecases
- Conclusion

Critical path usage

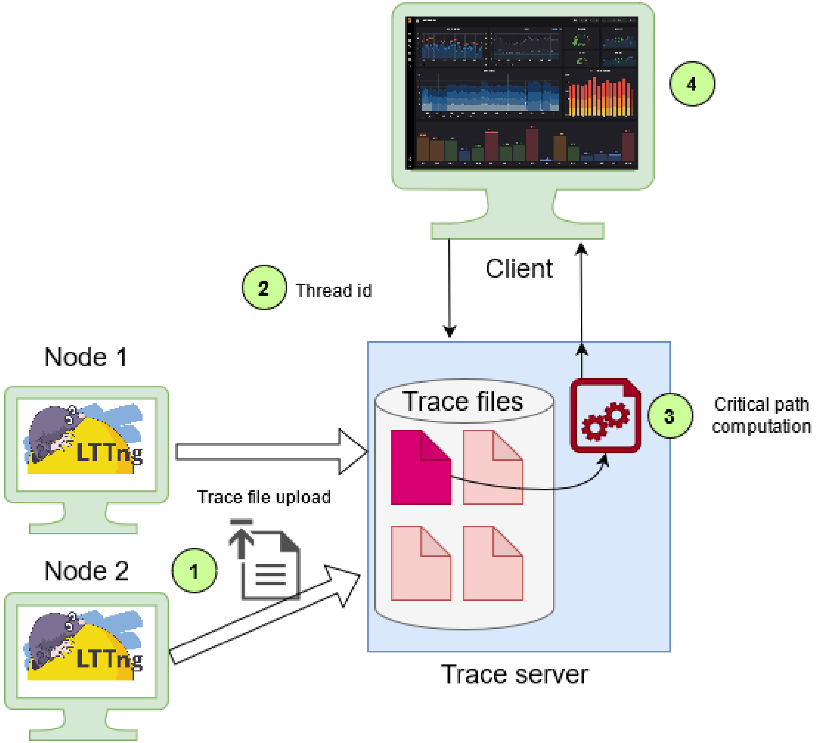
- Need for large distributed systems tracing
  - HPC systems
  - MPI clusters
  - Kubernetes and container clusters
- Transfer of trace files on analysis node was mandatory
- Critical path distributed computation was not optimized

“

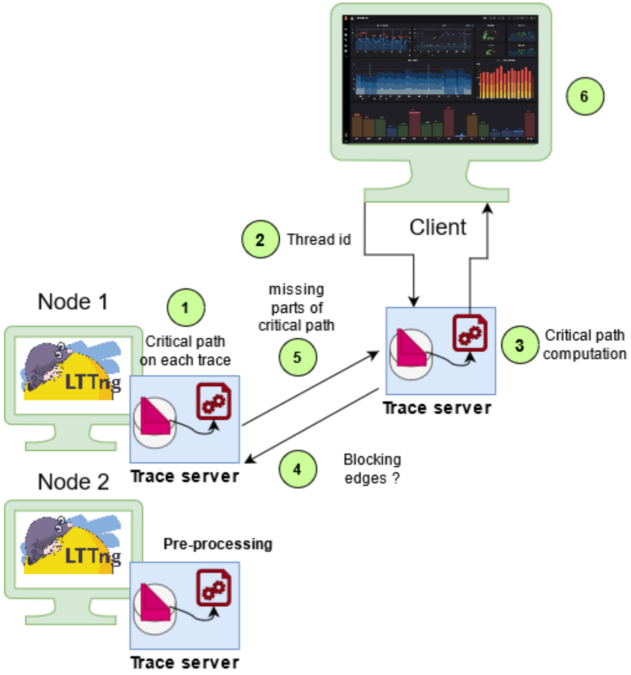
Critical path computation evolution

Actual architecture in Trace Compass (AL1)

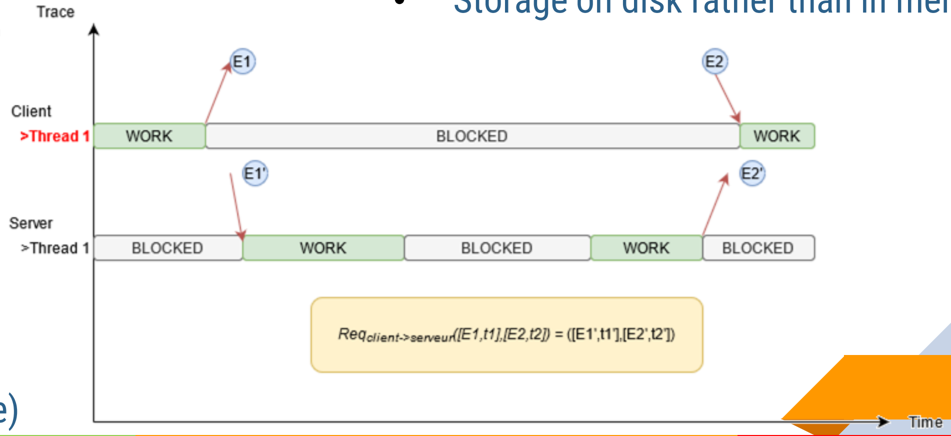
-



Parallelisation of the computation : algorithm (AL2)

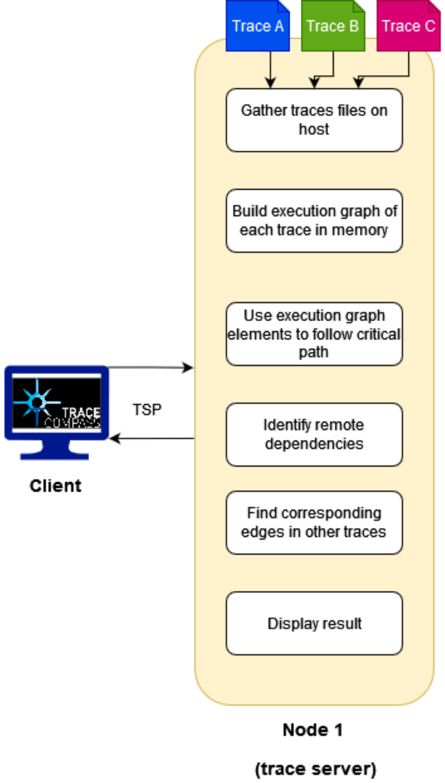


- Pre-processing of critical path on each node
- On client request, process the critical path of the trace, and ask only the missing parts of the path to other nodes
- Distributed processing, suitable for large number of nodes, less network load
- Storage on disk rather than in memory*

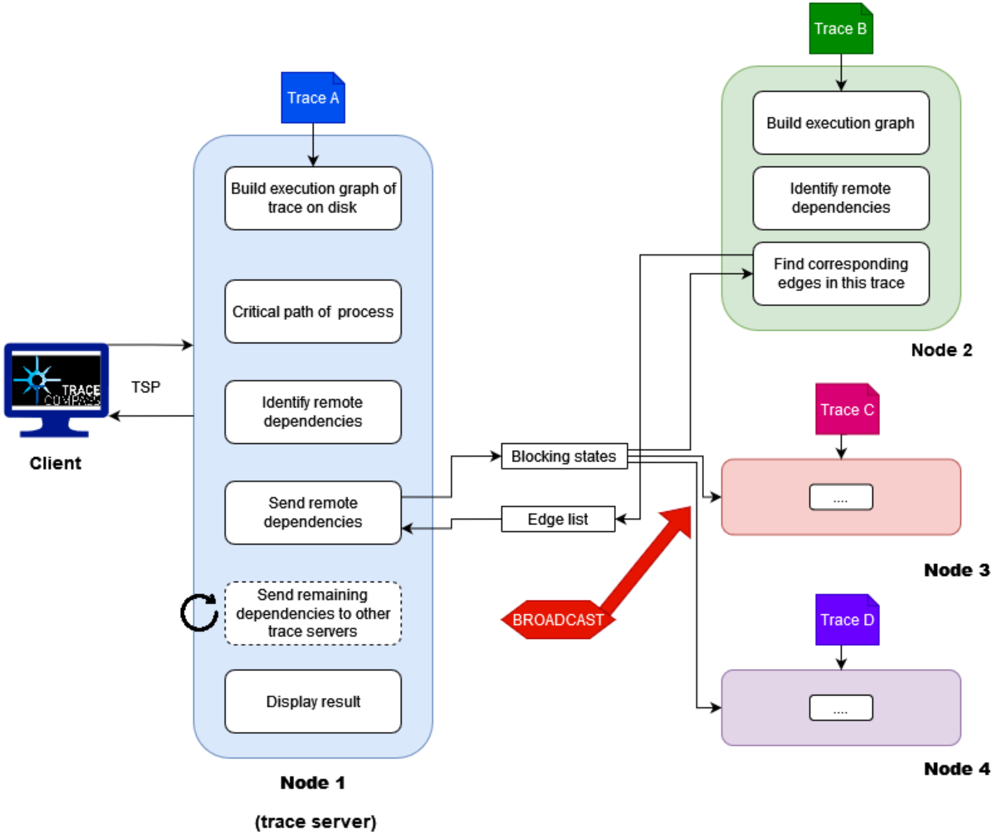


(*related work done by Arnaud and Geneviève)

AL1 vs AL2

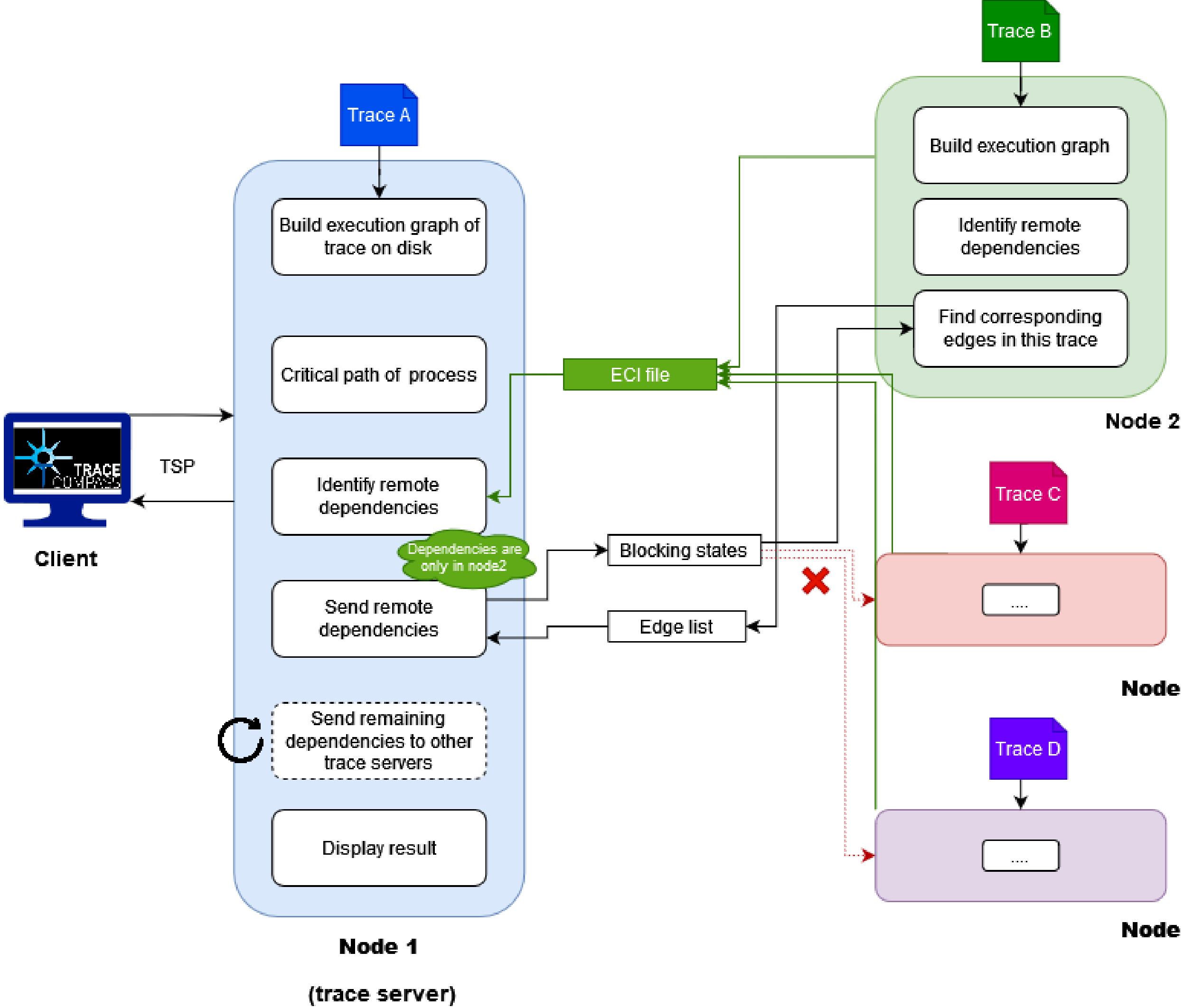

Trace A
Trace B
Trace C
Gather traces files on host
Build execution graph of each trace in memory
Use execution graph elements to follow critical path
Identify remote dependencies
Find corresponding edges in other traces
Display result
TSP
TRACE COMPASS
Client
Node 1
(trace server)

Actual algorithm (AL1)



Distributed version (AL3)

AL2 vs AL3


Trace A
Build execution graph of trace on disk
Critical path of process
Identify remote dependencies
Dependencies are only in node2
Send remote dependencies
Send remaining dependencies to other trace servers
Display result
Node 1
(trace server)
TRACE COMPASS
TSP
Client
ECI file
Blocking states
Edge list
Trace B
Build execution graph
Identify remote dependencies
Find corresponding edges in this trace
Node 2
Trace C
....
Node 3
Trace D
....
Node 4

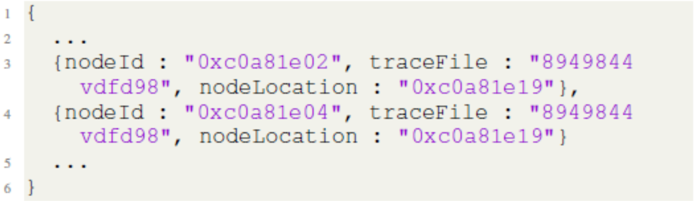
My work (AL 3) : External communication Index

- Improve algorithm to :

Introduce External communication Index (ECI) usage

- Index file exchanged after execution graph processing

Identify remote dependencies and location of remote trace



```
{
  ...
  {nodeId : "0xc0a81e02", traceFile : "8949844
    vdfd98", nodeLocation : "0xc0a81e19"},
  {nodeId : "0xc0a81e04", traceFile : "8949844
    vdfd98", nodeLocation : "0xc0a81e19"}
  ...
}
```

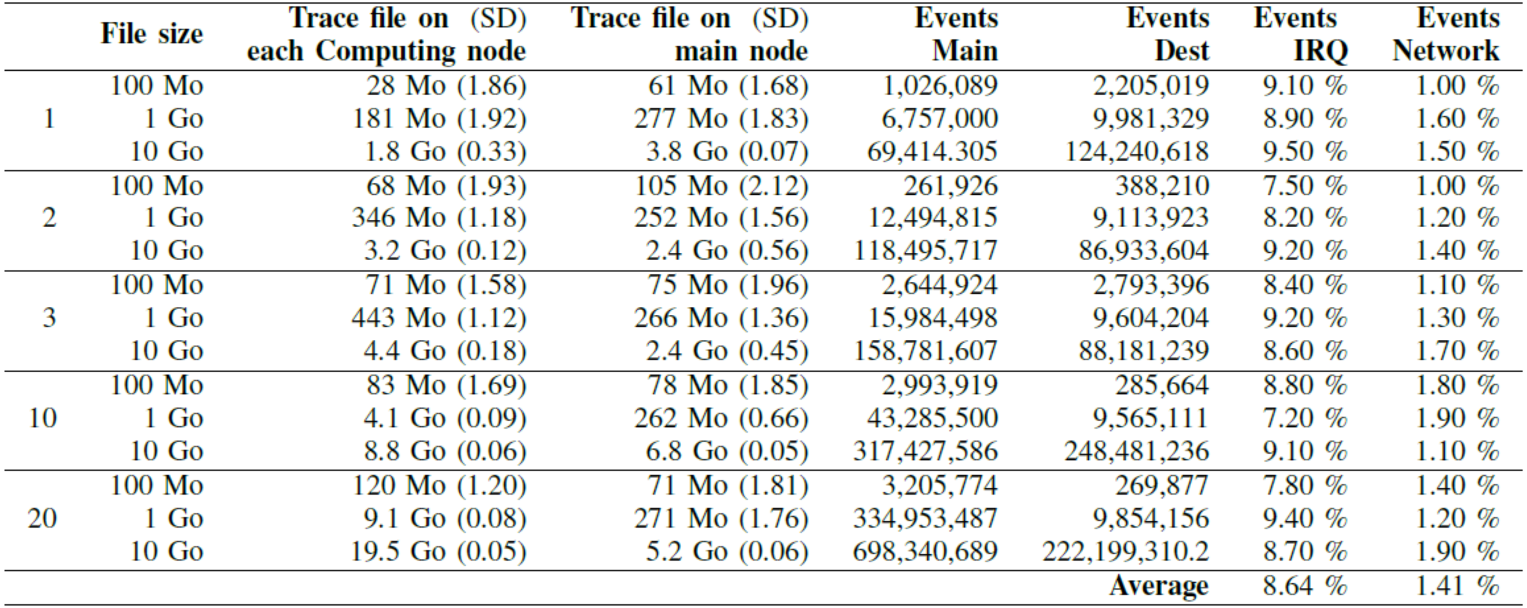
- Remove the need of broadcast communication on computing nodes to find remote dependencies

Pre-identification of remote dependencies for faster critical path processing

“

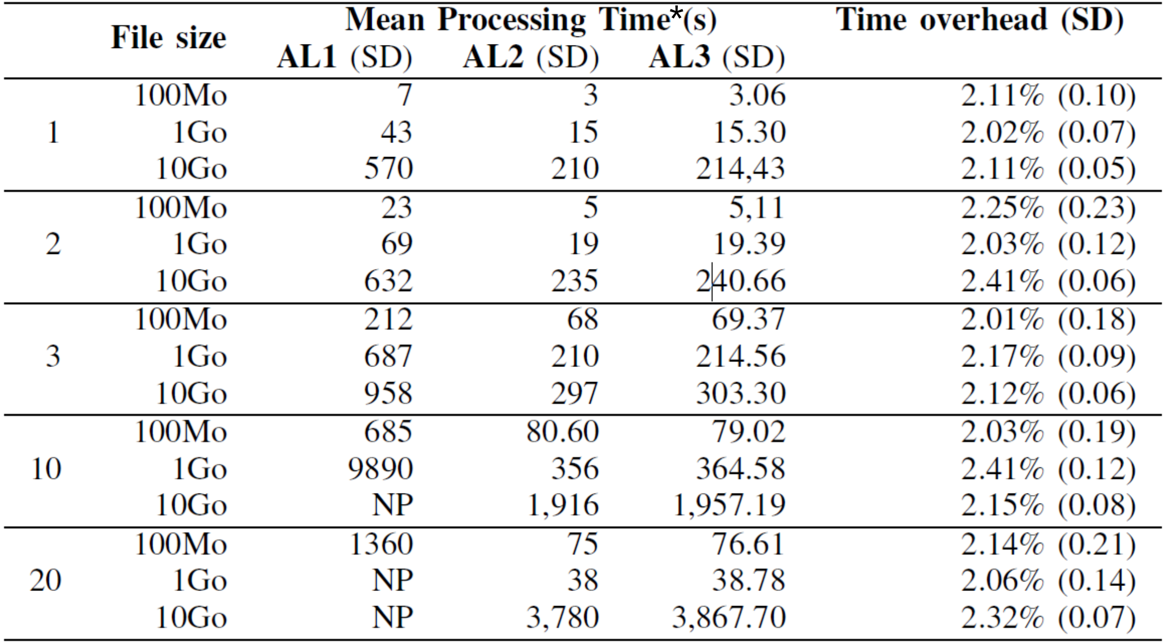
Benchmarks

SCP benchmark



	File size	Trace file on (SD) each Computing node	Trace file on (SD) main node	Events Main	Events Dest	Events IRQ	Events Network
1	100 Mo	28 Mo (1.86)	61 Mo (1.68)	1,026,089	2,205,019	9.10 %	1.00 %
	1 Go	181 Mo (1.92)	277 Mo (1.83)	6,757,000	9,981,329	8.90 %	1.60 %
	10 Go	1.8 Go (0.33)	3.8 Go (0.07)	69,414.305	124,240,618	9.50 %	1.50 %
2	100 Mo	68 Mo (1.93)	105 Mo (2.12)	261,926	388,210	7.50 %	1.00 %
	1 Go	346 Mo (1.18)	252 Mo (1.56)	12,494,815	9,113,923	8.20 %	1.20 %
	10 Go	3.2 Go (0.12)	2.4 Go (0.56)	118,495,717	86,933,604	9.20 %	1.40 %
3	100 Mo	71 Mo (1.58)	75 Mo (1.96)	2,644,924	2,793,396	8.40 %	1.10 %
	1 Go	443 Mo (1.12)	266 Mo (1.36)	15,984,498	9,604,204	9.20 %	1.30 %
	10 Go	4.4 Go (0.18)	2.4 Go (0.45)	158,781,607	88,181,239	8.60 %	1.70 %
10	100 Mo	83 Mo (1.69)	78 Mo (1.85)	2,993,919	285,664	8.80 %	1.80 %
	1 Go	4.1 Go (0.09)	262 Mo (0.66)	43,285,500	9,565,111	7.20 %	1.90 %
	10 Go	8.8 Go (0.06)	6.8 Go (0.05)	317,427,586	248,481,236	9.10 %	1.10 %
20	100 Mo	120 Mo (1.20)	71 Mo (1.81)	3,205,774	269,877	7.80 %	1.40 %
	1 Go	9.1 Go (0.08)	271 Mo (1.76)	334,953,487	9,854,156	9.40 %	1.20 %
	10 Go	19.5 Go (0.05)	5.2 Go (0.06)	698,340,689	222,199,310.2	8.70 %	1.90 %
					Average	8.64 %	1.41 %

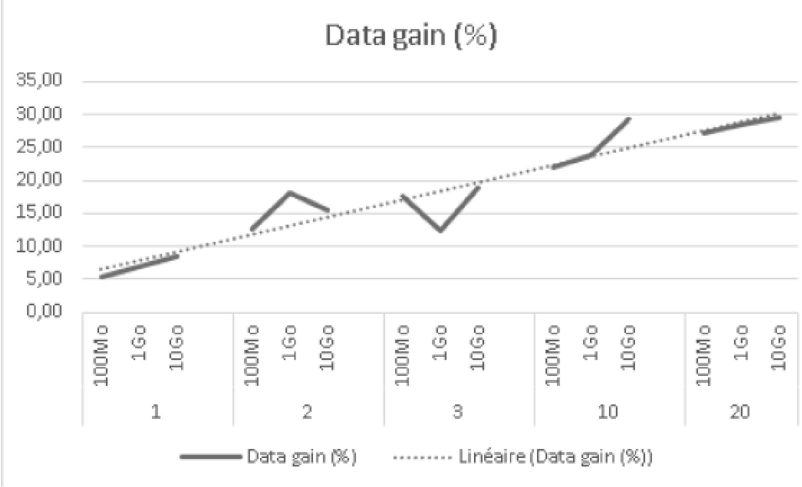
SCP benchmark

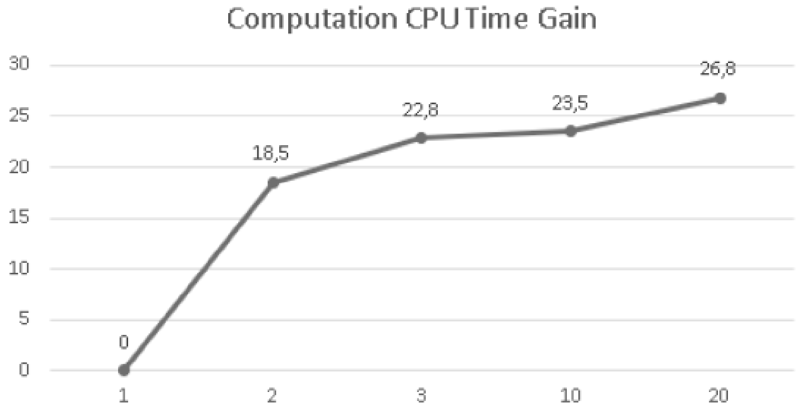


	File size	Mean Processing Time*(s)			Time overhead (SD)
		AL1 (SD)	AL2 (SD)	AL3 (SD)	
1	100Mo	7	3	3.06	2.11% (0.10)
	1Go	43	15	15.30	2.02% (0.07)
	10Go	570	210	214,43	2.11% (0.05)
2	100Mo	23	5	5,11	2.25% (0.23)
	1Go	69	19	19.39	2.03% (0.12)
	10Go	632	235	240.66	2.41% (0.06)
3	100Mo	212	68	69.37	2.01% (0.18)
	1Go	687	210	214.56	2.17% (0.09)
	10Go	958	297	303.30	2.12% (0.06)
10	100Mo	685	80.60	79.02	2.03% (0.19)
	1Go	9890	356	364.58	2.41% (0.12)
	10Go	NP	1,916	1,957.19	2.15% (0.08)
20	100Mo	1360	75	76.61	2.14% (0.21)
	1Go	NP	38	38.78	2.06% (0.14)
	10Go	NP	3,780	3,867.70	2.32% (0.07)

*Total CPU Time

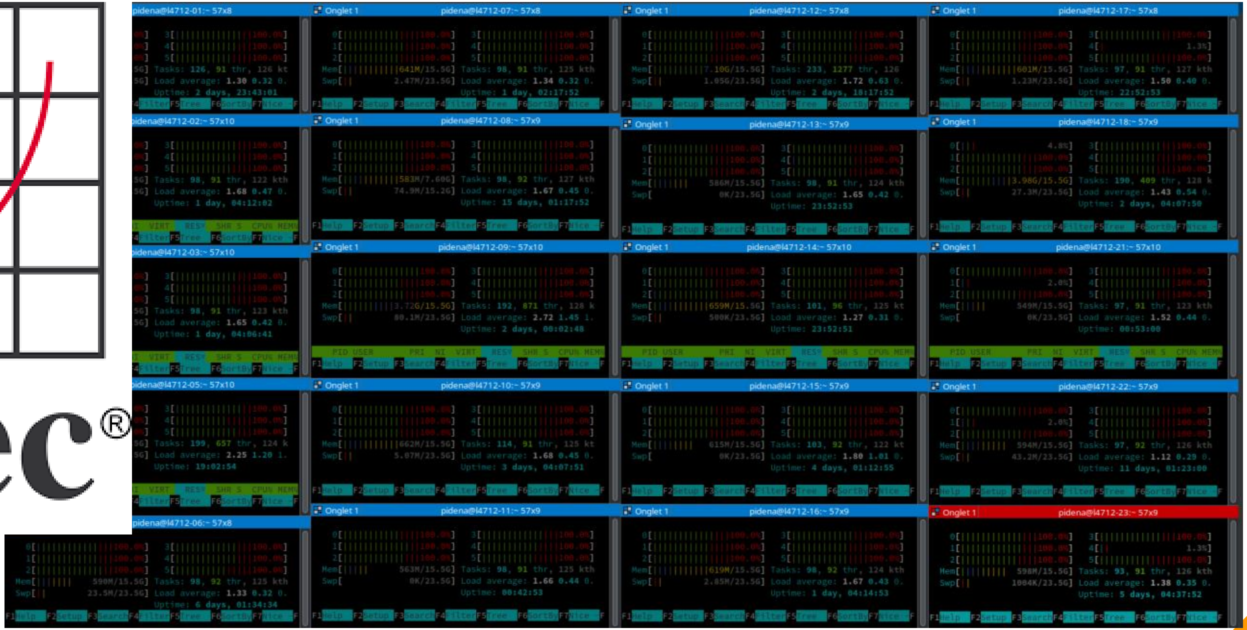
SCP benchmark





MPI benchmark

- 20 nodes with i5-9400 at 2.9Ghz with six cores and 16Go of memory



```
pidena@l4712-01:~ 57x8
0%]   3[|||||||||||||||||100.0%]
0%]   4[|||||||||||||||||100.0%]
0%]   5[|||||||||||||||||100.0%]
5G] Tasks: 126, 91 thr, 126 kt
5G] Load average: 1.30 0.32 0.
    Uptime: 2 days, 23:43:01
4FilterF5Tree  F6SortByF7Nice -F
```

```
pidena@l4712-02:~ 57x10
0%]   3[|||||||||||||||||100.0%]
0%]   4[|||||||||||||||||100.0%]
0%]   5[|||||||||||||||||100.0%]
5G] Tasks: 98, 91 thr, 122 kth
5G] Load average: 1.68 0.47 0.
    Uptime: 1 day, 04:12:02

I  VIRT   RES   SHR S  CPU% MEM%
4FilterF5Tree  F6SortByF7Nice -F
```

```
pidena@l4712-03:~ 57x10
0%]   3[|||||||||||||||||100.0%]
0%]   4[|||||||||||||||||100.0%]
0%]   5[|||||||||||||||||100.0%]
5G] Tasks: 98, 91 thr, 123 kth
5G] Load average: 1.65 0.42 0.
    Uptime: 1 day, 04:06:41

I  VIRT   RES   SHR S  CPU% MEM%
4FilterF5Tree  F6SortByF7Nice -F
```

```
pidena@l4712-05:~ 57x10
0%]   3[|||||||||||||||||100.0%]
0%]   4[|||||||||||||||||100.0%]
0%]   5[|||||||||||||||||100.0%]
5G] Tasks: 199, 657 thr, 124 k
5G] Load average: 2.25 1.20 1.
    Uptime: 19:02:54

I  VIRT   RES   SHR S  CPU% MEM%
4FilterF5Tree  F6SortByF7Nice -F
```

```
pidena@l4712-06:~ 57x8
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||||      590M/15.5G] Tasks: 98, 92 thr, 125 kth
 Swp[||          23.5M/23.5G] Load average: 1.33 0.32 0.
                              Uptime: 6 days, 01:34:34
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-07:~ 57x8
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||||||||641M/15.5G] Tasks: 98, 91 thr, 125 kth
 Swp[||          2.47M/23.5G] Load average: 1.34 0.32 0.
                              Uptime: 1 day, 02:17:52
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-08:~ 57x9
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||||||||583M/7.60G] Tasks: 98, 92 thr, 127 kth
 Swp[||          74.9M/15.2G] Load average: 1.67 0.45 0.
                              Uptime: 15 days, 01:17:52

F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-09:~ 57x10
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[||||||||||3.72G/15.5G] Tasks: 192, 871 thr, 128 k
 Swp[||          80.1M/23.5G] Load average: 2.72 1.45 1.
                              Uptime: 2 days, 00:02:48

   PID USER      PRI  NI  VIRT   RES   SHR S  CPU% MEM%
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-10:~ 57x9
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||||||||662M/15.5G] Tasks: 114, 91 thr, 125 kt
 Swp[||          5.07M/23.5G] Load average: 1.68 0.45 0.
                              Uptime: 3 days, 04:07:51

F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-11:~ 57x9
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||        563M/15.5G] Tasks: 98, 91 thr, 125 kth
 Swp[              0K/23.5G] Load average: 1.66 0.44 0.
                              Uptime: 00:42:53

F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-12:~ 57x8
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[||||||||||7.10G/15.5G] Tasks: 233, 1277 thr, 126
 Swp[||          1.05G/23.5G] Load average: 1.72 0.63 0.
                              Uptime: 2 days, 18:17:52
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-13:~ 57x9
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||||      586M/15.5G] Tasks: 98, 91 thr, 124 kth
 Swp[              0K/23.5G] Load average: 1.65 0.42 0.
                              Uptime: 23:52:53

F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-14:~ 57x10
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||||||||659M/15.5G] Tasks: 101, 96 thr, 125 kt
 Swp[||           500K/23.5G] Load average: 1.27 0.31 0.
                              Uptime: 23:52:51

   PID USER      PRI  NI  VIRT   RES   SHR S  CPU% MEM%
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-15:~ 57x9
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[||||||||     615M/15.5G] Tasks: 103, 92 thr, 122 kt
 Swp[              0K/23.5G] Load average: 1.80 1.01 0.
                              Uptime: 4 days, 01:12:55

F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-16:~ 57x9
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||||||||619M/15.5G] Tasks: 98, 92 thr, 124 kth
 Swp[||          2.85M/23.5G] Load average: 1.67 0.43 0.
                              Uptime: 1 day, 04:14:53

F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-17:~ 57x8
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|                  1.3%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||||||||601M/15.5G] Tasks: 97, 91 thr, 127 kth
 Swp[||          1.23M/23.5G] Load average: 1.50 0.40 0.
                              Uptime: 22:52:53
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-18:~ 57x9
   0[|||               4.8%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[||||||||||3.98G/15.5G] Tasks: 190, 409 thr, 128 k
 Swp[||          27.3M/23.5G] Load average: 1.43 0.54 0.
                              Uptime: 2 days, 04:07:50

F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

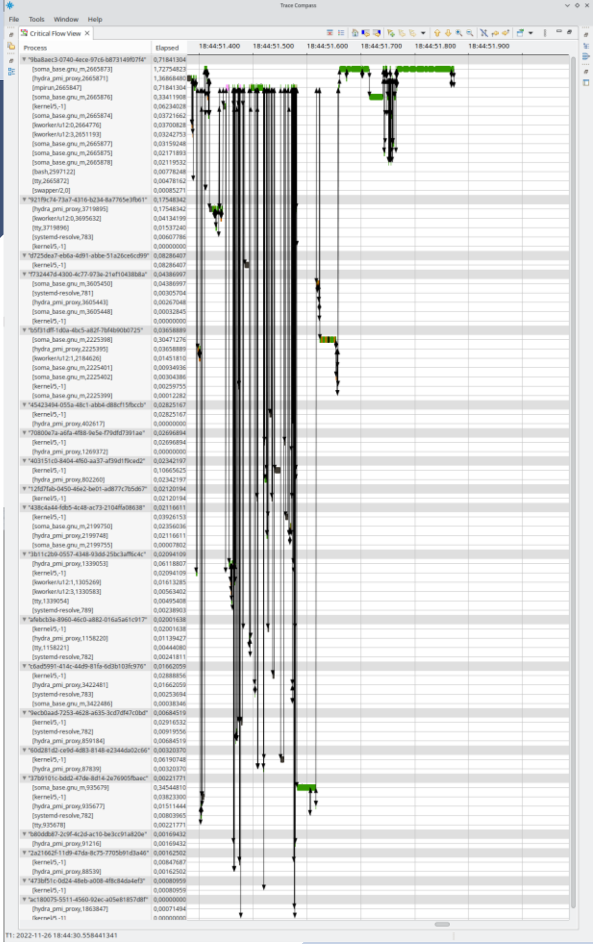
```
Onglet 1                  pidena@l4712-21:~ 57x10
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[||                2.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||        549M/15.5G] Tasks: 97, 91 thr, 123 kth
 Swp[              0K/23.5G] Load average: 1.52 0.44 0.
                              Uptime: 00:53:00

   PID USER      PRI  NI  VIRT   RES   SHR S  CPU% MEM%
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-22:~ 57x9
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||               2.0%]   4[|||||||||||||||||100.0%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||||||     594M/15.5G] Tasks: 97, 92 thr, 126 kth
 Swp[||          43.2M/23.5G] Load average: 1.12 0.29 0.
                              Uptime: 11 days, 01:23:00

F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

```
Onglet 1                  pidena@l4712-23:~ 57x9
   0[|||||||||||||||||100.0%]   3[|||||||||||||||||100.0%]
   1[|||||||||||||||||100.0%]   4[||                 1.3%]
   2[|||||||||||||||||100.0%]   5[|||||||||||||||||100.0%]
 Mem[|||||||||     598M/15.5G] Tasks: 93, 91 thr, 126 kth
 Swp[||          1004K/23.5G] Load average: 1.38 0.35 0.
                              Uptime: 5 days, 04:37:52

F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F
```

MPI benchmark

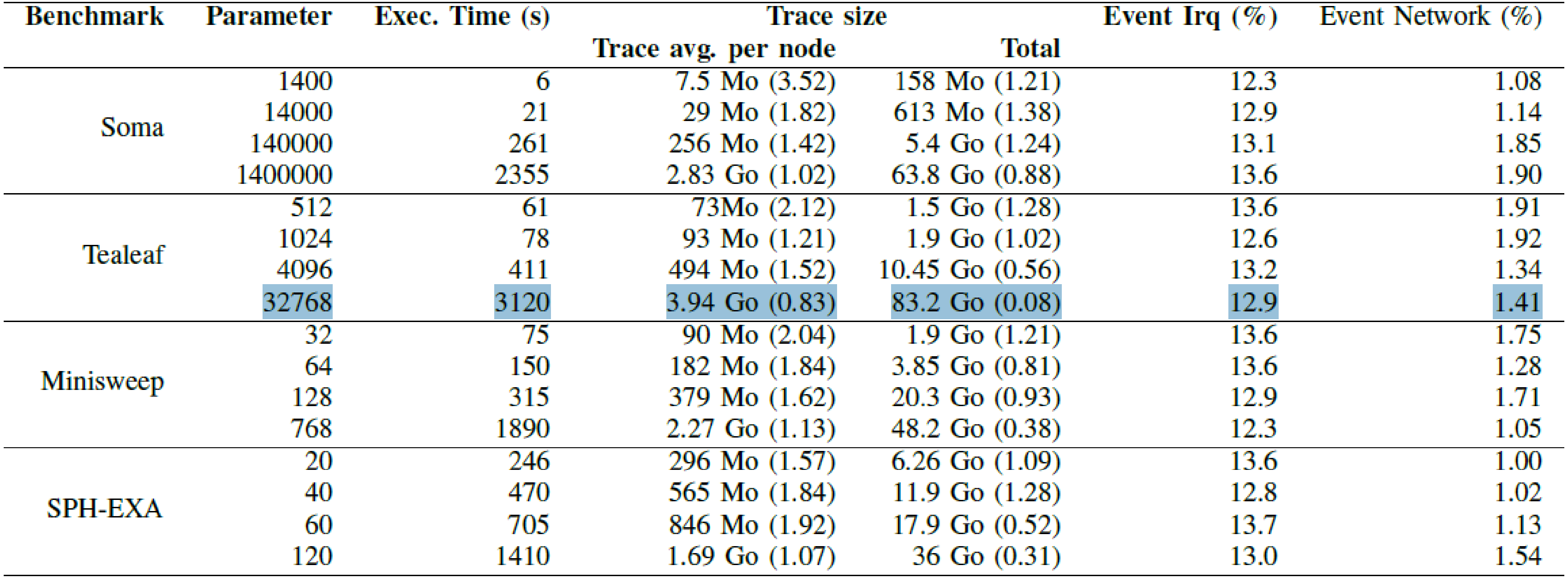
- SOMA : Offers Monte-Carlo Acceleration is a High-Performance Computing Monte-Carlo simulation for soft coarse-grained polymers. The variable load is the *number of simulated polymers*.
- Tealeaf : solves the linear heat conduction equation on a spatially decomposed regular grid. The variable load is the *number of cells*.
- Minisweep: is a Nuclear Engineering and radiation transport simulation. The variable load is the *grid cell size*.
- SPH-EXA: performs hydrodynamical and computational fluid dynamics simulations. The variable load is the *number of particles to the cube*.



Process	Elapsed
▼ "9ba8aec3-0740-4ece-97c6-b873149f07f4"	0,71841304
[soma_base.gnu_m,2665873]	1,72754823
[hydra_pmi_proxy,2665871]	1,36868480
[mpirun,2665847]	0,71841304
[soma_base.gnu_m,2665876]	0,33411908
[kernel/5,-1]	0,06234028
[soma_base.gnu_m,2665874]	0,03721662
[kworker/u12:0,2664776]	0,03700828
[kworker/u12:3,2651193]	0,03242753
[soma_base.gnu_m,2665877]	0,03159248
[soma_base.gnu_m,2665875]	0,02171893
[soma_base.gnu_m,2665878]	0,02119532
[bash,2597122]	0,00778248
[tty,2665872]	0,00478162
[swapper/2,0]	0,00085271
▼ "921f9c74-73a7-4316-b234-8a7765e3fb61"	0,17548342
[hydra_pmi_proxy,3719895]	0,17548342
[kworker/u12:0,3695632]	0,04134199
[tty,3719896]	0,01537240
[systemd-resolve,783]	0,00607786
[kernel/5,-1]	0,00000000
▼ "d725dea7-eb6a-4d91-abbe-51a26ce6cd99"	0,08286407
[kernel/5,-1]	0,08286407
▼ "f732447d-4300-4c77-973e-21ef10438b8a"	0,04386997
[soma_base.gnu_m,3605450]	0,04386997
[systemd-resolve,781]	0,00305704
[hydra_pmi_proxy,3605443]	0,00267048
[soma_base.gnu_m,3605448]	0,00032845
[kernel/5,-1]	0,00000000
▼ "b5f31dff-1d0a-4bc5-a82f-7bf4b90b0725"	0,03658889
[soma_base.gnu_m,2225398]	0,30471276
[hydra_pmi_proxy,2225395]	0,03658889
[kworker/u12:1,2184626]	0,01451810
[soma_base.gnu_m,2225401]	0,00934936
[soma_base.gnu_m,2225402]	0,00304386
[kernel/5,-1]	0,00259755
[soma_base.gnu_m,2225399]	0,00012282
▼ "45423494-055a-48c1-abb4-d88cf15fbccb"	0,02825167
[kernel/5,-1]	0,02825167
[hydra_pmi_proxy,402617]	0,00000000
▼ "70800e7a-a6fa-4f88-9e5e-f79dfd7391ae"	0,02696894
[kernel/5,-1]	0,02696894
[hydra_pmi_proxy,1269372]	0,00000000
▼ "403151c0-8404-4f60-aa37-af39d1f9ced2"	0,02342197
[kernel/5,-1]	0,10665625
[hydra_pmi_proxy,802260]	0,02342197
▼ "12fd7fab-0450-46e2-be01-ad877c7b5d67"	0,02120194
[kernel/5,-1]	0,02120194
▼ "438c4a44-fdb5-4c48-ac73-2104ffa08638"	0,02116611
[kernel/5,-1]	0,03926153
[soma_base.gnu_m,2199750]	0,02356036
[hydra_pmi_proxy,2199748]	0,02116611
[soma_base.gnu_m,2199755]	0,00007802
▼ "3b11c2b9-0557-4348-93dd-25bc3aff6c4c"	0,02094109
[hydra_pmi_proxy,1339053]	0,06118807
[kernel/5,-1]	0,02094109
[kworker/u12:1,1305269]	0,01613285
[kworker/u12:3,1330583]	0,00563402
[tty,1339054]	0,00495408
[systemd-resolve,789]	0,00238903
▼ "afebcb3e-8960-46c0-a882-016a5a61c917"	0,02001638
[kernel/5,-1]	0,02001638
[hydra_pmi_proxy,1158220]	0,01139427
[tty,1158221]	0,00444080
[systemd-resolve,782]	0,00241811
▼ "c6ad5991-414c-44d9-81fa-6d3b103fc976"	0,01662059
[kernel/5,-1]	0,02888856
[hydra_pmi_proxy,3422481]	0,01662059
[systemd-resolve,783]	0,00253694
[soma_base.gnu_m,3422486]	0,00038346
▼ "9ecb0aad-7253-4628-a635-3cd7df47c0bd"	0,00684519
[kernel/5,-1]	0,02916532
[systemd-resolve,782]	0,00919556
[hydra_pmi_proxy,859184]	0,00684519
▼ "60d281d2-ce9d-4d83-8148-e2344da02c66"	0,00320370
[kernel/5,-1]	0,06190748
[hydra_pmi_proxy,87839]	0,00320370
▼ "37b9101c-bdd2-47de-8d14-2e76905fbaec"	0,00221771
[soma_base.gnu_m,935679]	0,34544810
[kernel/5,-1]	0,03823300
[hydra_pmi_proxy,935677]	0,01511444
[systemd-resolve,782]	0,00803965
[tty,935678]	0,00221771
▼ "b80ddb87-2c9f-4c2d-ac10-be3cc91a820e"	0,00169432
[hydra_pmi_proxy,91216]	0,00169432
▼ "2a21662f-11d9-47da-8c75-7705b91d3a46"	0,00162502
[kernel/5,-1]	0,00847687
[hydra_pmi_proxy,88539]	0,00162502
▼ "473bf51c-0d24-48eb-a008-4f8c84da4ef3"	0,00080959
[kernel/5,-1]	0,00080959
▼ "ac180075-5511-4560-92ec-a05e81857d8f"	0,00000000
[hydra_pmi_proxy,1863847]	0,00071494
[kernel/5,-1]	0,00000000

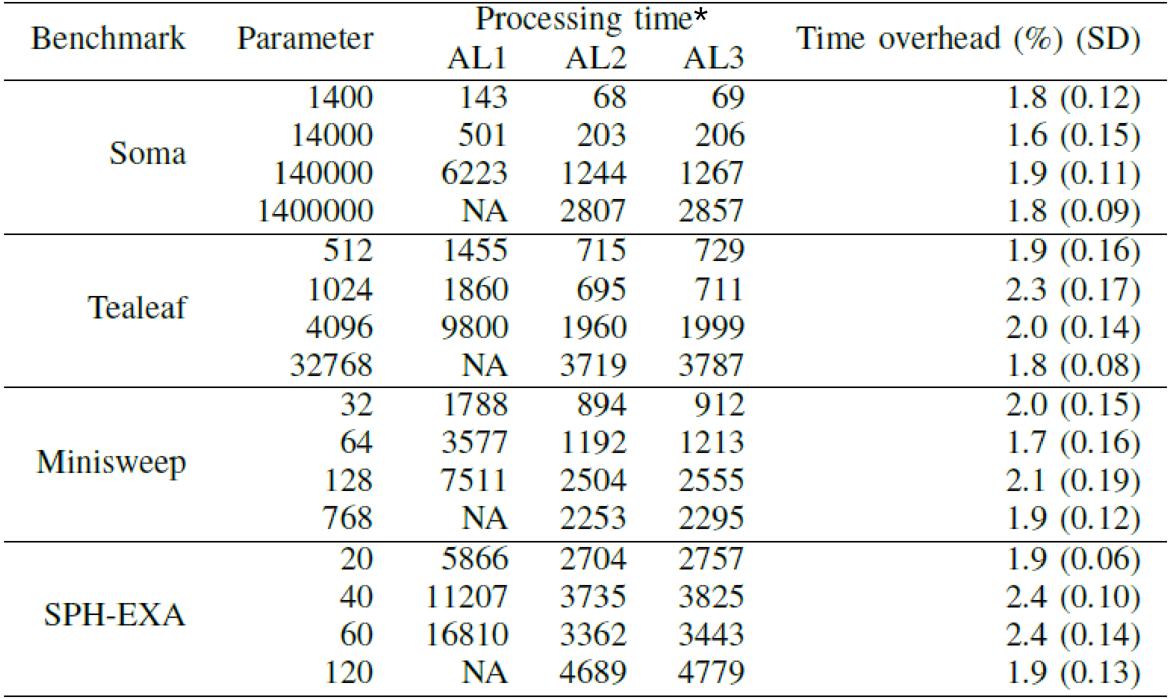
T1: 2022-11-26 18:44:30.558441341

MPI benchmark



Benchmark	Parameter	Exec. Time (s)	Trace size		Event Irq (%)	Event Network (%)
			Trace avg. per node	Total		
Soma	1400	6	7.5 Mo (3.52)	158 Mo (1.21)	12.3	1.08
	14000	21	29 Mo (1.82)	613 Mo (1.38)	12.9	1.14
	140000	261	256 Mo (1.42)	5.4 Go (1.24)	13.1	1.85
	1400000	2355	2.83 Go (1.02)	63.8 Go (0.88)	13.6	1.90
Tealeaf	512	61	73Mo (2.12)	1.5 Go (1.28)	13.6	1.91
	1024	78	93 Mo (1.21)	1.9 Go (1.02)	12.6	1.92
	4096	411	494 Mo (1.52)	10.45 Go (0.56)	13.2	1.34
	32768	3120	3.94 Go (0.83)	83.2 Go (0.08)	12.9	1.41
Minisweep	32	75	90 Mo (2.04)	1.9 Go (1.21)	13.6	1.75
	64	150	182 Mo (1.84)	3.85 Go (0.81)	13.6	1.28
	128	315	379 Mo (1.62)	20.3 Go (0.93)	12.9	1.71
	768	1890	2.27 Go (1.13)	48.2 Go (0.38)	12.3	1.05
SPH-EXA	20	246	296 Mo (1.57)	6.26 Go (1.09)	13.6	1.00
	40	470	565 Mo (1.84)	11.9 Go (1.28)	12.8	1.02
	60	705	846 Mo (1.92)	17.9 Go (0.52)	13.7	1.13
	120	1410	1.69 Go (1.07)	36 Go (0.31)	13.0	1.54

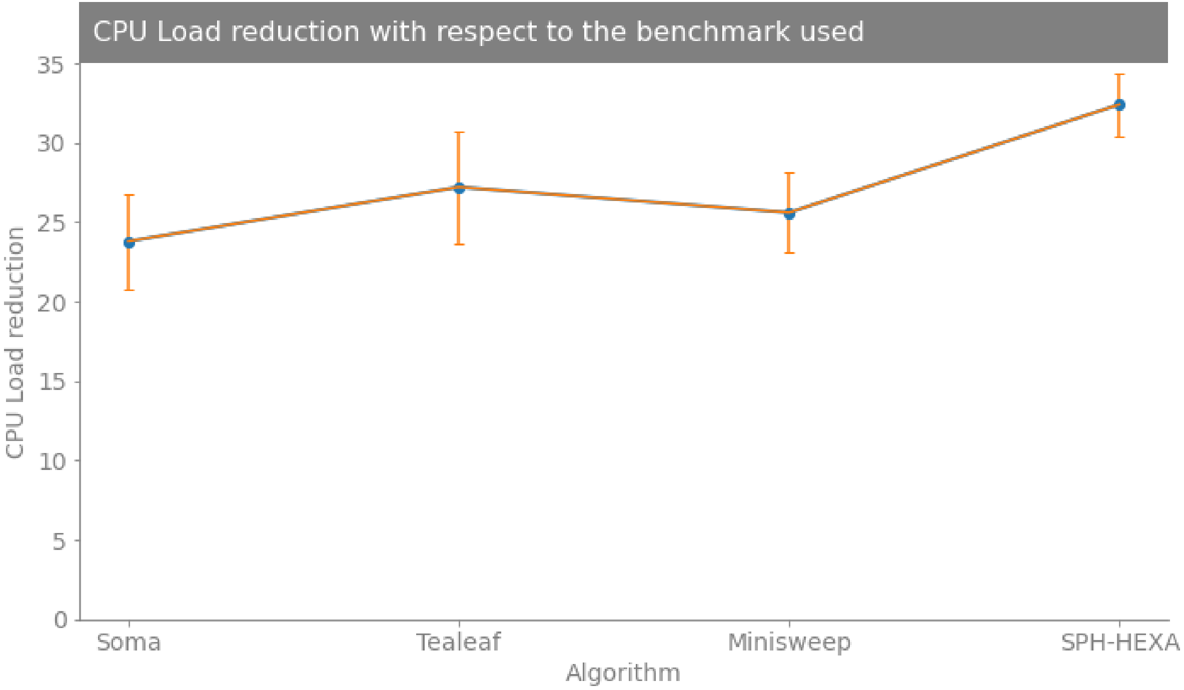
MPI benchmark



Benchmark	Parameter	Processing time*			Time overhead (%) (SD)
		AL1	AL2	AL3	
Soma	1400	143	68	69	1.8 (0.12)
	14000	501	203	206	1.6 (0.15)
	140000	6223	1244	1267	1.9 (0.11)
	1400000	NA	2807	2857	1.8 (0.09)
Tealeaf	512	1455	715	729	1.9 (0.16)
	1024	1860	695	711	2.3 (0.17)
	4096	9800	1960	1999	2.0 (0.14)
	32768	NA	3719	3787	1.8 (0.08)
Minisweep	32	1788	894	912	2.0 (0.15)
	64	3577	1192	1213	1.7 (0.16)
	128	7511	2504	2555	2.1 (0.19)
	768	NA	2253	2295	1.9 (0.12)
SPH-EXA	20	5866	2704	2757	1.9 (0.06)
	40	11207	3735	3825	2.4 (0.10)
	60	16810	3362	3443	2.4 (0.14)
	120	NA	4689	4779	1.9 (0.13)

*Total CPU Time

MPI benchmark



“

Future work and usecases

What remains to be done ?

- Remote time synchronisation of traces
- Better protocol for graph elements exchanges
- Automatic coordination between nodes

Usecases

- Target usecases :
  - MPI cluster : follow a MPI task between computing nodes
  - Kubernetes cluster : follow a request in a distributed web application
  - ZeroMQ communication : follow a message exchange between several containers

“

Conclusion

Conclusion

- Improvement and benchmark of critical path Distributed computation
- Next step : Integration of Critical path in Trace Server Protocol (for Theia and Grafana viewers)
- Extend parallelisation to other kind of analysis

Thank you for listening !