Spatio-Temporal Aggregation of StarPU multi-node traces

Lucas Mello Schnorr, Lucas Assis
Instituto de Informática, UFRGS

– NumPex / ExaSoft / WP5 –
(chez Datamove)
June 19th, 2025

Introduction & Motivation

StarPU-MPI: Task Programming over Clusters of Machines Enhanced with Accelerators: https://inria.hal.science/hal-00725477

- Each node generates a FXT file with timestamped events
- Voluminous traces with all application tasks (and many other data)
  - Start/End of states, performance metrics

Introduction & Motivation

StarPU-MPI: Task Programming over Clusters of Machines Enhanced with Accelerators: https://inria.hal.science/hal-00725477

- Each node generates a FXT file with timestamped events
- Voluminous traces with all application tasks (and many other data)
  - Start/End of states, performance metrics

Assumptions

1. Provide an exploratory analysis of the application/runtime behavior
   - We don't know possible performance issues → minimal filtering during tracing
   - Justify performance problems with contextual information from traces
2. Trace visualization overwhelmed by the amount of data (temporal/spatial)
   - Necessity of a visualization → Visualizing More Performance Data Than What Fits on Your Screen: https://inria.hal.science/hal-00737651

Introduction & Motivation

StarPU-MPI: Task Programming over Clusters of Machines Enhanced with Accelerators: https://inria.hal.science/hal-00725477

- Each node generates a FXT file with timestamped events
- Voluminous traces with all application tasks (and many other data)
  - Start/End of states, performance metrics

Assumptions

1. Provide an exploratory analysis of the application/runtime behavior
   - We don't know possible performance issues → minimal filtering during tracing
   - Justify performance problems with contextual information from traces
2. Trace visualization overwhelmed by the amount of data (temporal/spatial)
   - Necessity of a visualization → Visualizing More Performance Data Than What Fits on Your Screen: https://inria.hal.science/hal-00737651

Problems

(1) Too much but necessary traces; (2) Visualization scalability of space/time views

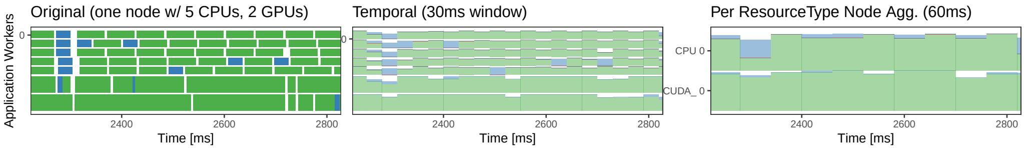
Objetive & Approach

1. Do trace aggregation before the trace visualization
2. Investigate specifically the spatial/temporal aggregation together
3. Provide a traditional space/time view

Objetive & Approach

1. Do trace aggregation before the trace visualization
2. Investigate specifically the spatial/temporal aggregation together
3. Provide a traditional space/time view

Existing efforts within StarVZ (https://cran.r-project.org/web/packages/starvz/)



Original (one node w/ 5 CPUs, 2 GPUs)

Temporal (30ms window)

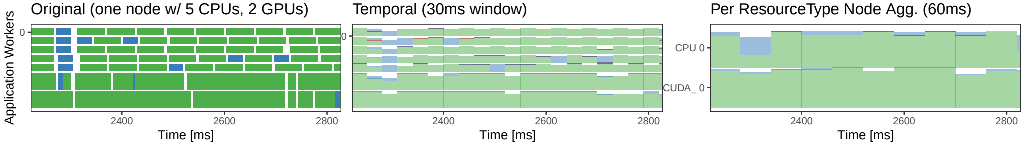
Per ResourceType Node Agg. (60ms)

→ It lacks spatial aggregation (i. e. aggregate nodes with similar behavior)

Objetive & Approach

1. Do trace aggregation before the trace visualization
2. Investigate specifically the spatial/temporal aggregation together
3. Provide a traditional space/time view

Existing efforts within StarVZ (https://cran.r-project.org/web/packages/starvz/)



→ It lacks spatial aggregation (i. e. aggregate nodes with similar behavior)

Goal

Explore `lpaggreg` within the context of StarVZ for StarPU traces | Dosimont et. al. "A spatiotemporal data aggregation technique for performance analysis of large-scale execution traces". CLUSTER 2014 → https://github.com/dosimont/lpaggreg

Methodology & Workflow

(A) In the cluster

1. Run the experiment in the cluster
2. Collect FXT traces
3. Run StarVZ Phase 1 script → PARQUET (Columnar-based files)

(B) In the laptop

1. Employ StarVZ R Package
2. Integrate lpaggreg (this work)
   - `read_starvz`, export pjdump, micro, aggregation, viz

Methodology & Workflow

(A) In the cluster

1. Run the experiment in the cluster
2. Collect FXT traces
3. Run StarVZ Phase 1 script → PARQUET (Columnar-based files)

(B) In the laptop

1. Employ StarVZ R Package
2. Integrate lpaggreg (this work)
    - `read_starvz`, export pjdump, micro, aggregation, viz

(C) Evaluation

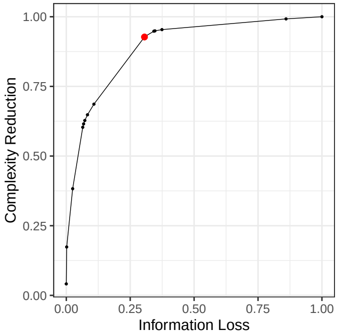
- Use previous StarPU traces obtained with (A)
    - Nesi et. al. "Summarizing task-based applications behavior over many nodes through progression clustering". PDP 2023. | Chameleon + ExaGeoStat
- Stress the usage of lpaggreg features (Information Loss, Complexity Reduction)

Lpaggreg features (Information Loss, Complexity Reduction)

Integrate lpaggreg (this work)

- `read_starvz`, export pjdump, micro, aggregation, viz

Parameter	Gain	Loss	POpt
0	0.0411956541727119	-1.02641907782919e-16	FALSE
0.0625	0.173655727924065	0.00157655769097006	FALSE
0.125	0.382826967688981	0.0249004552694742	FALSE
0.1875	0.603343589271184	0.0644300849930824	FALSE
0.25	0.615363211728487	0.0677042920117212	FALSE
0.3125	0.627709273385034	0.0724654636070262	FALSE
0.375	0.648086722611672	0.0828704306131623	FALSE
0.4375	0.686311544823401	0.108093646729556	FALSE
0.5	0.927290907242805	0.305838193382574	TRUE
0.5625	0.929053986692192	0.307990944412131	FALSE
0.625	0.93092906631553	0.311047076192813	FALSE
0.6875	0.948104037152438	0.342731084421385	FALSE
0.8125	0.949305929084811	0.34671405810594	FALSE
0.875	0.953750338312438	0.374181737848363	FALSE
0.9375	0.99223795542665	0.859582936312456	FALSE
1	1	1	FALSE



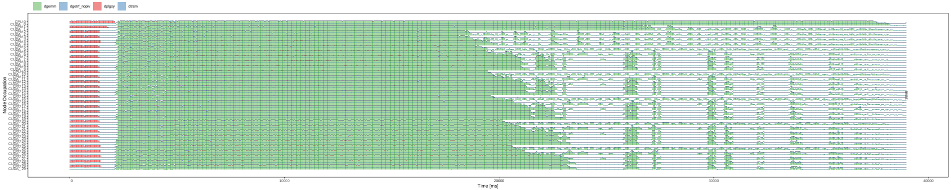
Each spatio/temporal aggregation provides several views

- One for each Parameter: 0 means minimal aggregation; 1 means full aggregation
- Each parameter represent a tradeoff (with an "ideal tradeoff", see POpt)

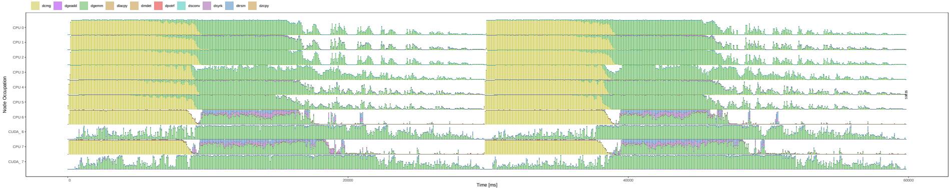
Case studies (Chameleon Dense LU Facto. and ExaGeoStat)

2W+DIF: 30 nodes, 2 GPUs each, two faulty nodes with only one GPU each

- It uses StarPU-Simgrid (http://dx.doi.org/10.1002/cpe.3555) to run Chameleon

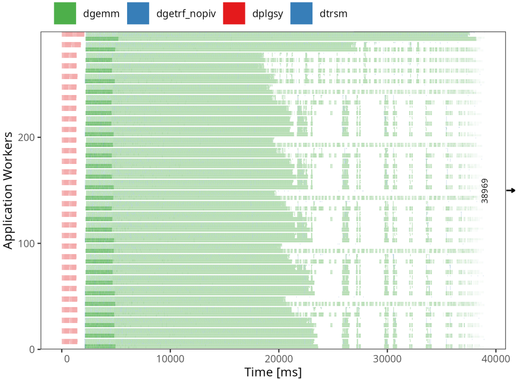


EXAGEO: 8 nodes, where six are CPU-only (2iters; real execution of ExaGeoStat in G5K)

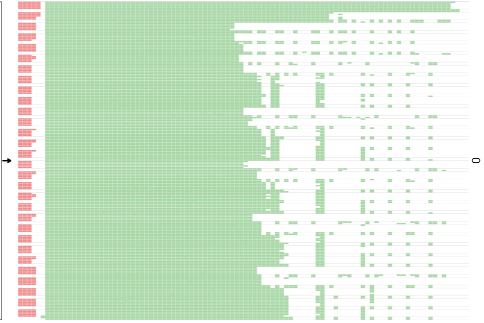


Preliminary Results: Dense LU Facto 1/2

StarVZ (no aggregation)



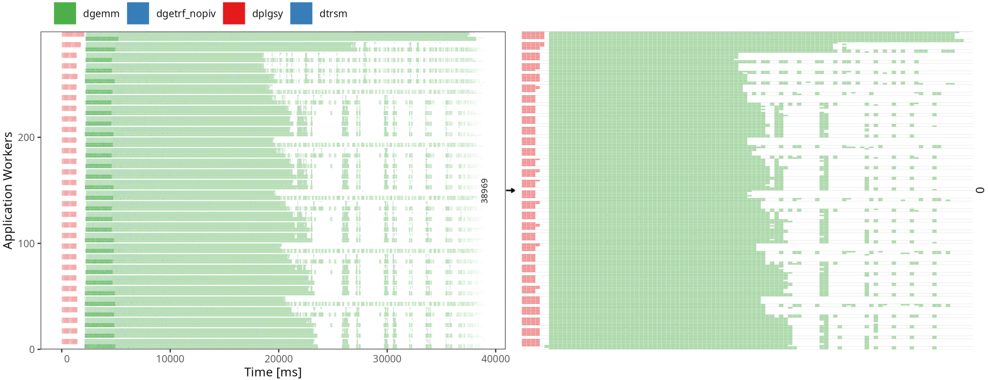
Lpaggreg Viz (minimal aggregation)
100 time slices



Preliminary Results: Dense LU Facto 1/2

StarVZ (no aggregation)

Lpaggreg Viz (minimal aggregation)
100 time slices

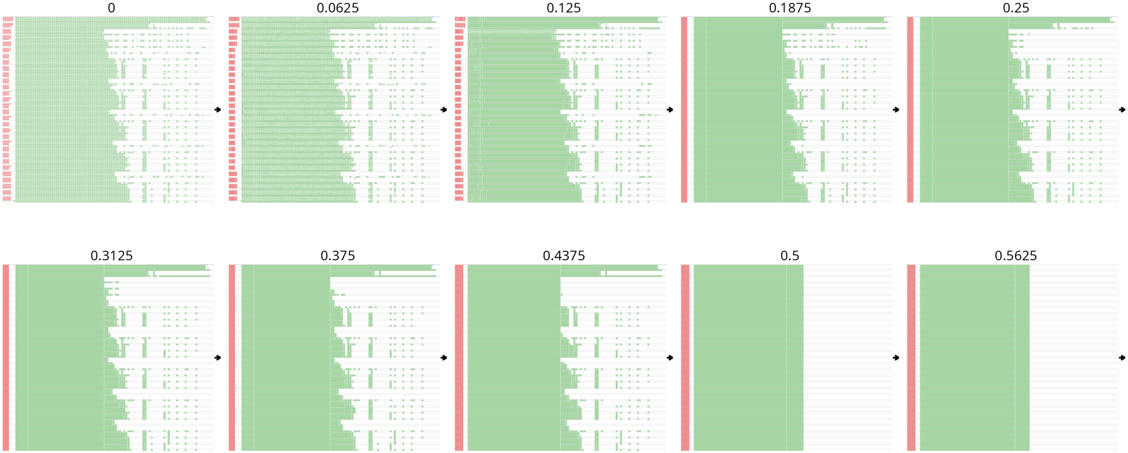


Overall visualization simplification (much less graphical elements)

- "minimal aggregation" may be more detailed with more time slices

Preliminary Results: Dense LU Facto 2/2

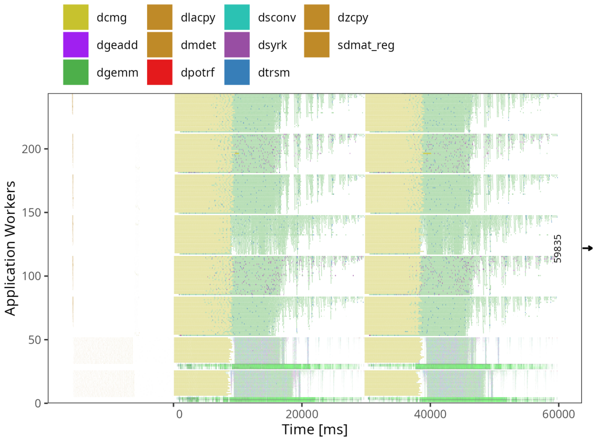
The first 10 tradeoffs (0.5 is POpt)

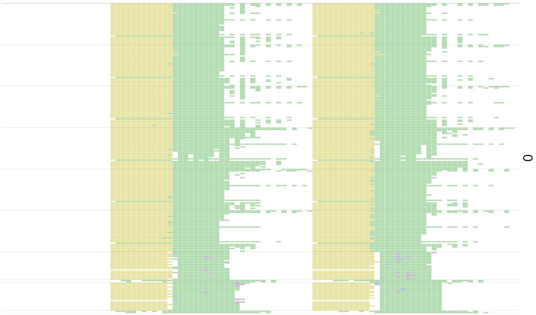


Preliminary Results: ExaGeoStat 1/2

StarVZ (no aggregation)

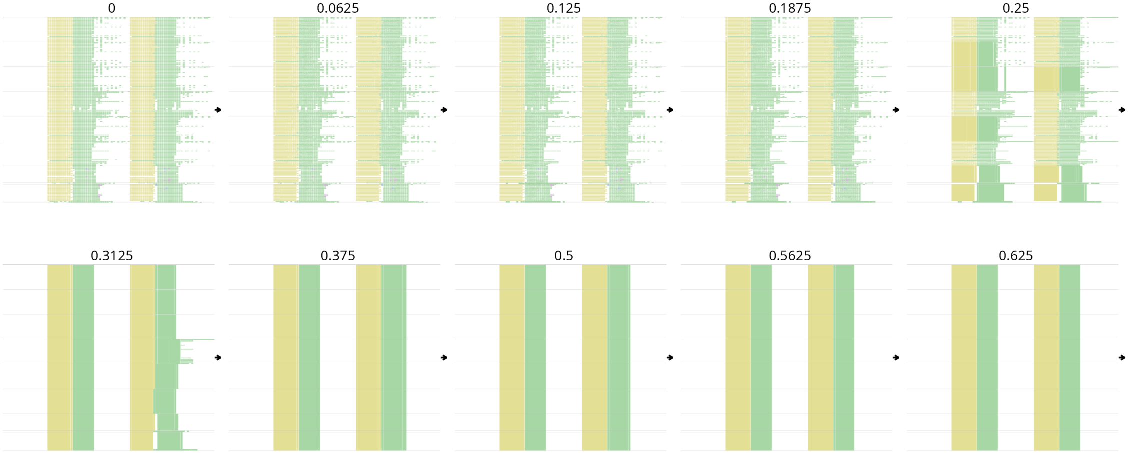
Lpaggreg Viz (minimal aggregation)
100 time slices





Preliminary Results: ExaGeoStat 2/2

The first 10 tradeoffs (0.375 is POpt)



Conclusion & Future Work

Weaknesses

- Computationally expensive (not linear) as the number of time slices increases
  - But we can run this part in the cluster anyway
- Method adopts a single and flat hierarchy
  - Spatial aggregation only works with neighbor nodes

Conclusion & Future Work

Weaknesses

- Computationally expensive (not linear) as the number of time slices increases
  - But we can run this part in the cluster anyway
- Method adopts a single and flat hierarchy
  - Spatial aggregation only works with neighbor nodes

Multiple hierarchies

- Nesi et. al. calculates a “Progression Metric” for each node (PDP23)
  - Multiple node groups with similar progressions
- Incorporate these groups into a lpaggreg hierarchy
  - With intermediate levels enriching the "flat" version of today
  - Slicing the trace and using several different well-selected hierarchies
    - Surely a visualization challenge! ;-)

Conclusion & Future Work

Weaknesses

- Computationally expensive (not linear) as the number of time slices increases
    - But we can run this part in the cluster anyway
- Method adopts a single and flat hierarchy
    - Spatial aggregation only works with neighbor nodes

Multiple hierarchies

- Nesi et. al. calculates a “Progression Metric” for each node (PDP23)
    - Multiple node groups with similar progressions
- Incorporate these groups into a lpaggreg hierarchy
    - With intermediate levels enriching the "flat" version of today
    - Slicing the trace and using several different well-selected hierarchies
        - Surely a visualization challenge! ;-)

Machinery is working. Do some realistic performance analysis with it.

- Experimentation, What-if scenarios, etc → Large-scale experiments

Contact

Merci pour votre attention !

Lucas Mello Schnorr <schnorr@inf.ufrgs.br>
Lucas Barros de Assis <lucas.assis@inf.ufrgs.br>