Visualization of Parallel or Distributed Systems

Jean-Marc Vincent and Arnaud Legrand

Laboratory ID-IMAG
MESCAL Project
Universities of Grenoble
{Jean-Marc.Vincent,Arnaud.Legrand}@imag.fr

November 13, 2011

Outline

A good survey on this topic

Lucas Schnorr's PhD thesis: *Some Visualization Models applied to the Analysis of Parallel Applications*

http://tel.archives-ouvertes.fr/tel-00459443/en/

Outline

A First Illustrative Example

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of X = 9.0
Mean of Y = 7.5

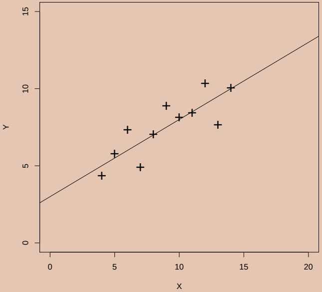
Correlation = 0.816

A First Illustrative Example

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Res. stdev = 1.237
Correlation = 0.816

Scatter plot

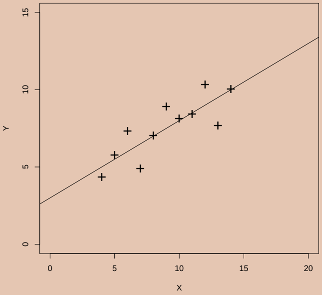


A First Illustrative Example

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Res. stdev = 1.237
Correlation = 0.816

Scatter plot



1. The data set "behaves like" a linear curve with some scatter;
2. There is no justification for a more complicated model (e.g., quadratic);
3. There are no outliers;
4. The vertical spread of the data appears to be of equal height irrespective of the X-value; this indicates that the data are equally-precise throughout and so a "regular" (that is, equi-weighted) fit is appropriate.

A First Illustrative Example

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Res. stdev = 1.237
Correlation = 0.816

$X^{(2)}$	$Y^{(2)}$
10.00	9.14
8.00	8.14
13.00	8.74
9.00	8.77
11.00	9.26
14.00	8.10
6.00	6.13
4.00	3.10
12.00	9.13
7.00	7.26
5.00	4.74

$N = 11$ samples
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Res. stdev = 1.237
Correlation = 0.816

$X^{(3)}$	$Y^{(3)}$
10.00	7.46
8.00	6.77
13.00	12.74
9.00	7.11
11.00	7.81
14.00	8.84
6.00	6.08
4.00	5.39
12.00	8.15
7.00	6.42
5.00	5.73

$N = 11$ samples
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Res. stdev = 1.237
Correlation = 0.816

$X^{(4)}$	$Y^{(4)}$
8.00	6.58
8.00	5.76
8.00	7.71
8.00	8.84
8.00	8.47
8.00	7.04
8.00	5.25
19.00	12.50
8.00	5.56
8.00	7.91
8.00	6.89

$N = 11$ samples
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Res. stdev = 1.237
Correlation = 0.816

A First Illustrative Example

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Res. stdev = 1.237
Correlation = 0.816

$X^{(2)}$	$Y^{(2)}$
10.00	9.14

Res. stdev = 1.237
Correlation = 0.816

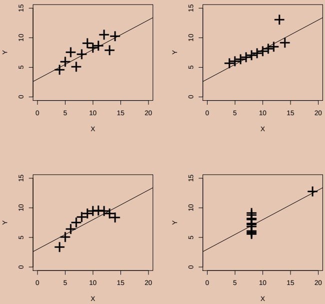
$X^{(3)}$	$Y^{(3)}$
10.00	7.46

Res. stdev = 1.237
Correlation = 0.816

$X^{(4)}$	$Y^{(4)}$
8.00	6.58

Res. stdev = 1.237
Correlation = 0.816

Scatter plot



A First Illustrative Example

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(2)}$	$Y^{(2)}$
10.00	9.14

Res. stdev $= 1.237$
Correlation $= 0.816$

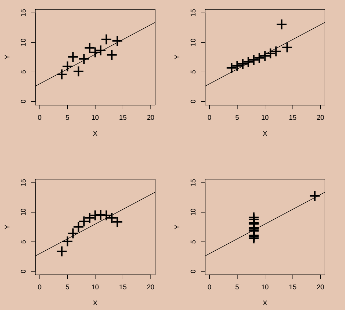
$X^{(3)}$	$Y^{(3)}$
10.00	7.46

Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(4)}$	$Y^{(4)}$
8.00	6.58

Res. stdev $= 1.237$
Correlation $= 0.816$

Scatter plot



1. data set 1 is clearly linear with some scatter.
2. data set 2 is clearly quadratic.
3. data set 3 clearly has an outlier.
4. data set 4 is obviously the victim of a poor experimental design with a single point far removed from the bulk of the data "wagging the dog".

Problem statement

- All analysis we perform rely on (sometimes implicit) assumptions. If these assumptions do not hold, the analysis will be a complete non-sense.
- Checking these assumptions is not always easy and sometimes, it may even be difficult to list all these assumptions and formally state them.

A visualization can help to check these assumptions.

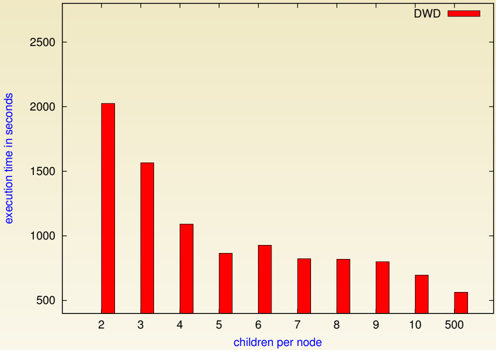
- Visual representation resort to our cognitive faculties to check properties.
  The visualization is meant to let us detect expected and unexpected behavior with respect to a given model.
- The problem is to represent on a limited space, typically a screen with a fixed resolution, a meaningful information about the behavior of an application or system.

Visualization and intuition

- Visualization can also be used to guide your intuition.
  Sometimes, you do not know exactly what you are looking for and looking at the data just helps.
- Some techniques (Exploratory Data Analysis) even build on this and propose to summarize main characteristics in easy-to-understand form, often with visual graphs, without using a statistical model or having formulated a hypothesis.
- Use with care, since visualization always have underlying models and when visualization is not adapted, what you may observe may be meaningless.
  Such approaches may help formulating hypothesis but these hypothesis have then to be tested upon new data-sets.
- Every visualization emphasizes some characteristics and hides others.
  Being aware of the underlying models helps choosing the right representation.

A "simple" graphical check for investigating speedup/scalability

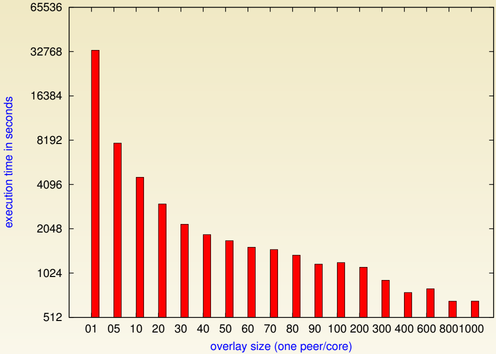
Plotting T_p versus p.



- y-axis does not start at 0, which makes speedup look more impressive
- x-axis is linear with an outlier.

A "simple" graphical check for investigating speedup/scalability

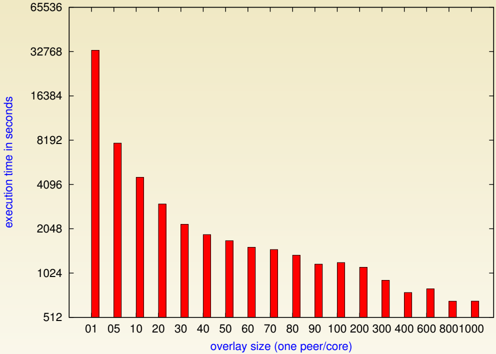
Plotting T_p versus p.



- y-axis uses log-scale
- x-axis is neither linear nor logarithmic so we cannot reason about the shape of the curve

A "simple" graphical check for investigating speedup/scalability

Plotting T_p versus p.

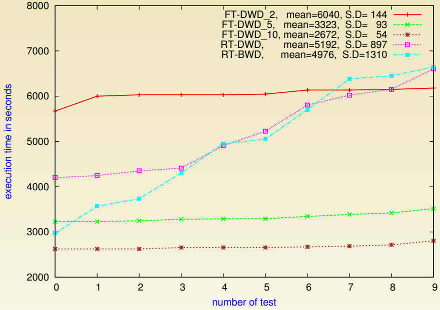


- y-axis uses log-scale
- x-axis is neither linear nor logarithmic so we cannot reason about the shape of the curve

Say, we want to test for Amhdal's law. Propose a better representation.

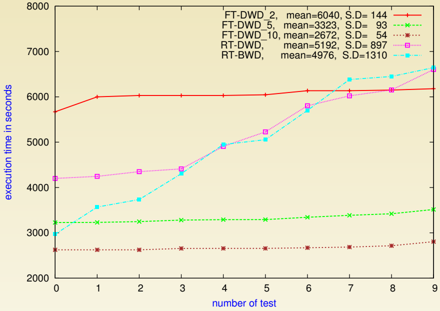
Graphically checking which alternative is better ?

5 different alternatives (FT-DWD_2, FT-DWD_5, FT-DWD_10, RT-DWD, RT-BWD), each tested 10 times.



Graphically checking which alternative is better ?

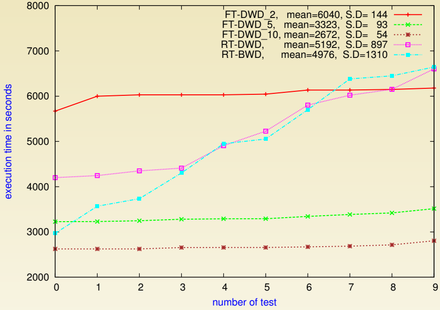
5 different alternatives (FT-DWD_2, FT-DWD_5, FT-DWD_10, RT-DWD, RT-BWD), each tested 10 times.



- Outcomes have been sorted by increasing value for each alternative and are then linked together.
- The line does not make any sense.
- Experiment order does not make any sense and makes it look like alternatives have been evaluated in 10 different settings (, which means they can be compared with each others for each setting).

Graphically checking which alternative is better ?

5 different alternatives (FT-DWD_2, FT-DWD_5, FT-DWD_10, RT-DWD, RT-BWD), each tested 10 times.



- Outcomes have been sorted by increasing value for each alternative and are then linked together.
- The line does not make any sense.
- Experiment order does not make any sense and makes it look like alternatives have been evaluated in 10 different settings (, which means they can be compared with each others for each setting).

Propose a better representation.

Read the basics

- For all such kind of “general” graphs where you summarize the results of several experiments, the very least you need to read is Jain’s book.
- It has check lists for “Good graphics”
- It presents the most common pitfalls in data representation
- It will teach how to cheat with your figures…
- …and how to detect cheaters. ;)

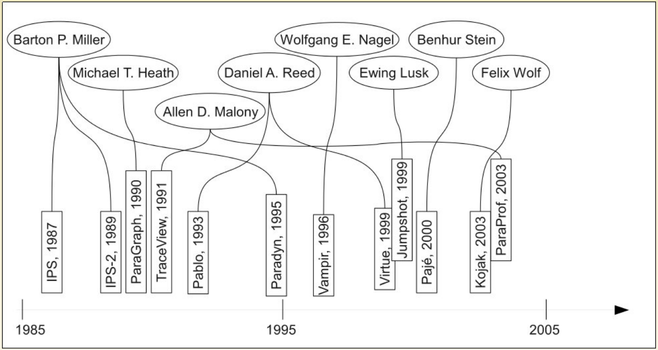
Outline

1. Data Representation
2. Visualization of Parallel System
   - Gantt Charts and Variations
   - Communication Matrix
   - Resource-oriented visualization
   - Treemaps
   - Statistical Charts
   - Fancy plots
   - Trace Comparison

Problem statement

- Assume we have a trace of parallel program.
- Traces are a series of time-stamped information about the states of the entities.
  It is a function:
  $$Entities \times Time \rightarrow State$$
  One is typically interested by its structure. E.g.,
  - time spent in a given state
  - state evolution
  - state "movement" or "propogation"
- Traces are a series of time-stamped communications, i.e.
  It is a function:
  $$(Entities \times Entities) \times Time \rightarrow (Informations \cup \{\emptyset\})$$
  Again, one is typically interested by its structure. E.g.,
  - amount of communication for a given category
  - communication structure
  - relation with entity state

Many tools



- Few are actually usable.
- Often follows the life-cycle of grants...
- Still, it doesn't mean it allows you to reinvent the wheel!

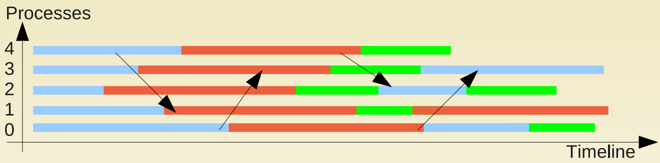
Outline

1. Data Representation

2. Visualization of Parallel System
   - Gantt Charts and Variations
   - Communication Matrix
   - Resource-oriented visualization
   - Treemaps
   - Statistical Charts
   - Fancy plots
   - Trace Comparison

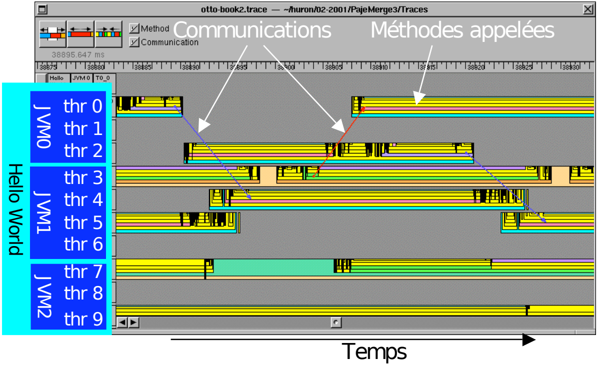
Gantt Chart

- A visualization technique created more than 100 years ago
- Initially used to organize and schedule the tasks of projects

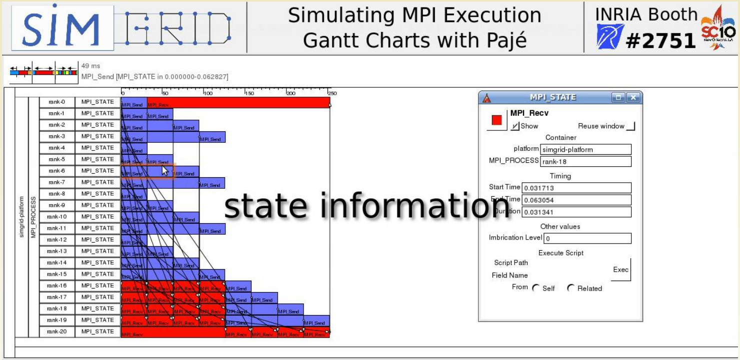


- This a "Spacetime" representation where colors are used to depict states. It enables to detect the critical path.
- Additional information often attached to rectangles (file name, line number, etc.).
- Main drawbacks:
  - Computer screens are limited in term of resolution
  - Scalability (needs scrollable and zoomable view)
  - Locality (entities organized along a single dimension)

Gantt Chart examples



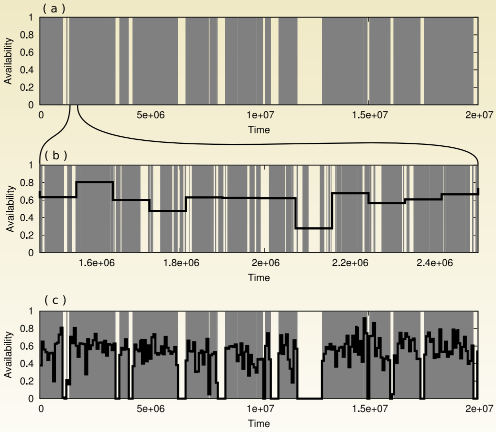
Gantt Chart examples



NAS DT Benchmark, Class A (BH), 21 processes

Visualizing a large number of information

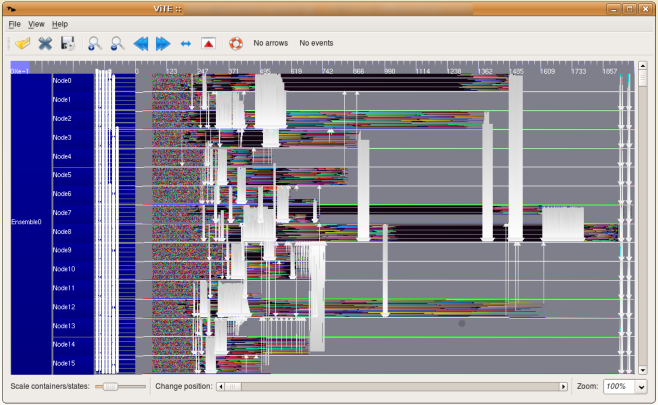
Many visualization tools perform this task either by overdrawing data onto the screen, thus eliminating parts of the information.
At best, they average values, thus smoothing everything. Yet this smoothing is often not what one would expect.



Time aggregation

Visualizing a large number of information

Many visualization tools perform this task either by overdrawing data onto the screen, thus eliminating parts of the information.
At best, they average values, thus smoothing everything. Yet this smoothing is often not what one would expect.



Spatial aggregation

Visualizing a large number of information

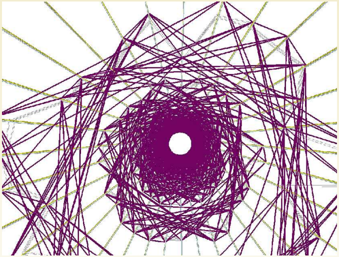
Many visualization tools perform this task either by overdrawing data onto the screen, thus eliminating parts of the information.
At best, they average values, thus smoothing everything. Yet this smoothing is often not what one would expect.

Warning

Although most tools do not do it:

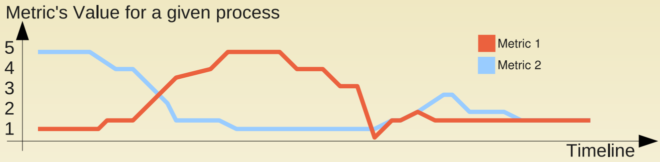
- Any aggregation or data reduction need to be advertised for.
- One should not rely on graphical aggregation operators (e.g. anti-aliasing) but use semantic ones.

When views do not use one dimension for time, the information need either to be aggregated on the whole duration or animated.
Depending on the chosen view, it may be more or less cumbersome. For Gantt charts, it could be tunnel views:



Variables in two or three dimensions

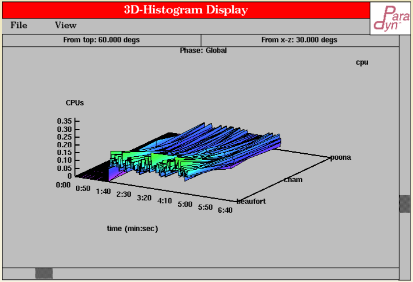
When state is not discrete or is too large and does not map well to colors, one need to resort to more classical plots.



- Typically, one select a set of entities and a set of variables to display. Such 2D graphs can be stacked or transformed into a 3D graph.
- Can be usefull to detect trends, correlations between variables or entities, . . .
- Again, this scales rather bad becomes cryptic when using more than a few variables and a dozen of entities

Variables in two or three dimensions

When state is not discrete or is too large and does not map well to colors, one need to resort to more classical plots.

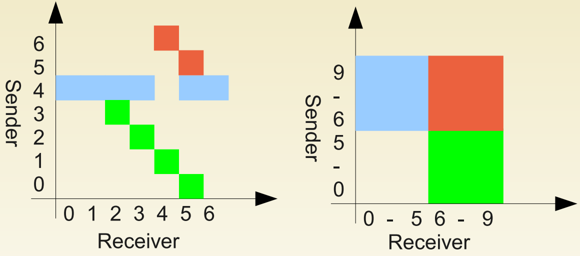


- Typically, one select a set of entities and a set of variables to display. Such 2D graphs can be stacked or transformed into a 3D graph.
- Can be usefull to detect trends, correlations between variables or entities, . . .
- Again, this scales rather bad becomes cryptic when using more than a few variables and a dozen of entities

Outline

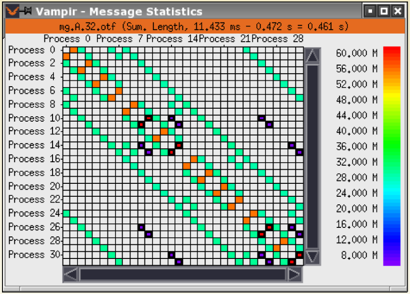
Communication Matrix

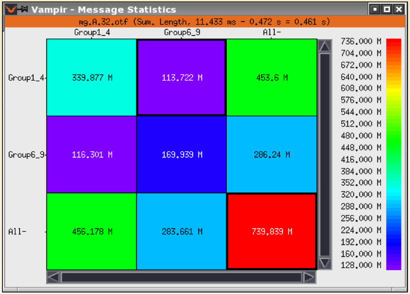
- For each pair of entities, compute statistics on the corresponding communications during a given time frame
- Colors can also be used to show additional information, such as the type of the communication, if it is collective or not, or the amount of the data transmitted.



- Time is not present in the view and hence, this view represents either instantaneous or aggregated values. The view can be animated over time.
- Main drawbacks:
  - Limited scalability although it may be possible to aggregated entities
  - Locality (entities organized along a single dimension)

Communication matrix examples

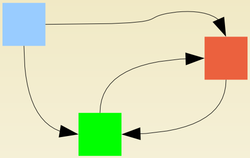




Outline

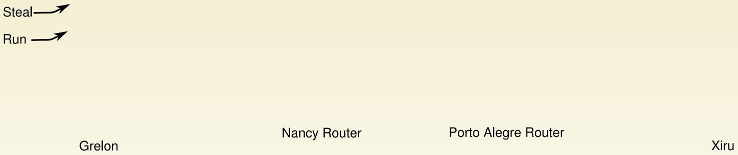
Platform graph

- Locality could be important to check that communications match the underlying topology.

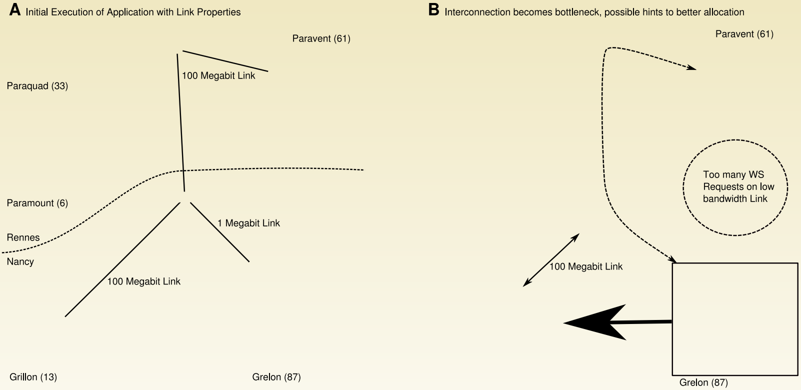


- Some tools provide "classical" topologies like rings, 3D-meshes or hypercubes.
- Time is not present in the view and hence, this view represents either instantaneous or aggregated values. The view can be animated over time.
- Main drawbacks:
  - Limited scalability

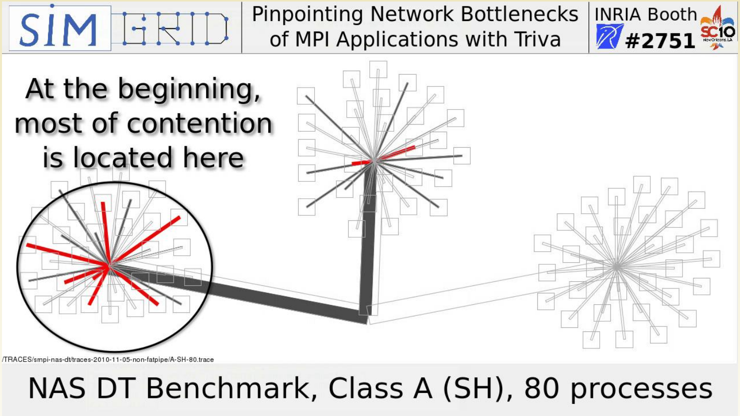
Platform-centric visualization



Platform-centric visualization



Platform-centric visualization



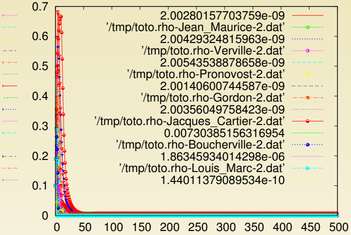
Setup

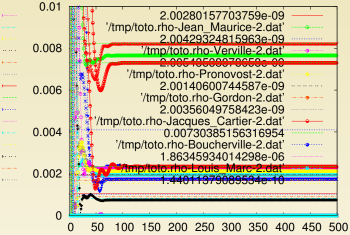
- 5 applications (with a specific color each) try to fairly share communication and computation resources.
- Different requirements in term of communication/computation
- Different origins
- Each resource adapts its price based on usage.
- Each application adapts its usage based on the price it has to pay.

Difficulties

- Spatial and temporal evolution
- Lots of variables/informations to visualize
- Scale issues (small/large values)

Setup

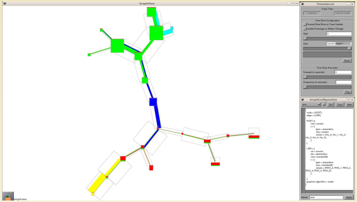


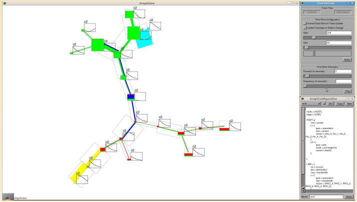


Difficulties

- Spatial and temporal evolution
- Lots of variables/informations to visualize
- Scale issues (small/large values)

Setup

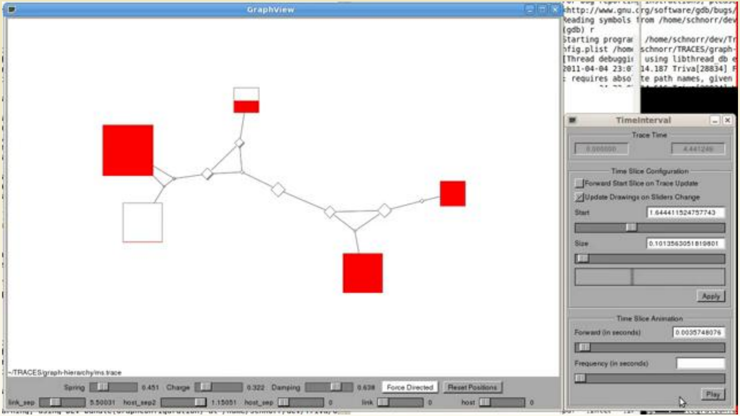




Difficulties

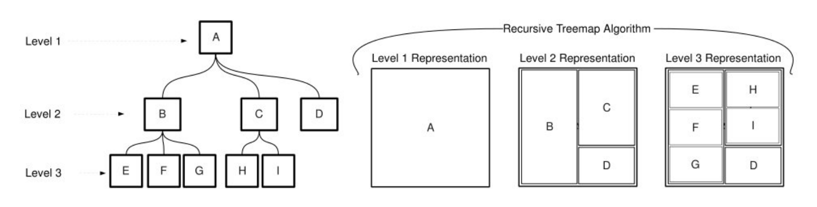
- Spatial and temporal evolution
- Lots of variables/informations to visualize
- Scale issues (small/large values)

Aggregation operations on graphs ?



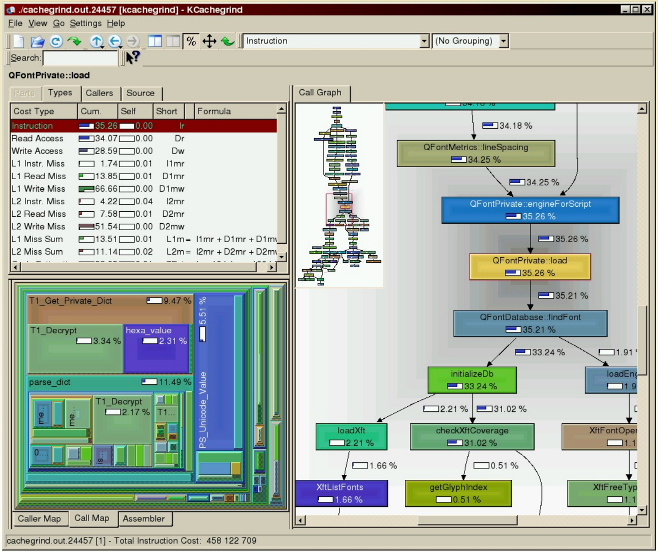
Outline

Treemaps

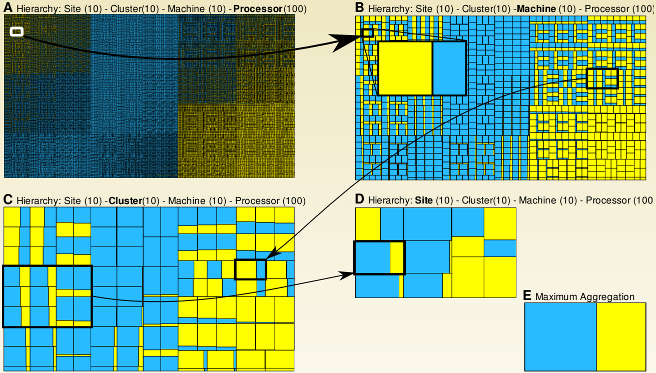


Level 1
Level 2
Level 3
A
B
C
D
E
F
G
H
I
Recursive Treemap Algorithm
Level 1 Representation
Level 2 Representation
Level 3 Representation

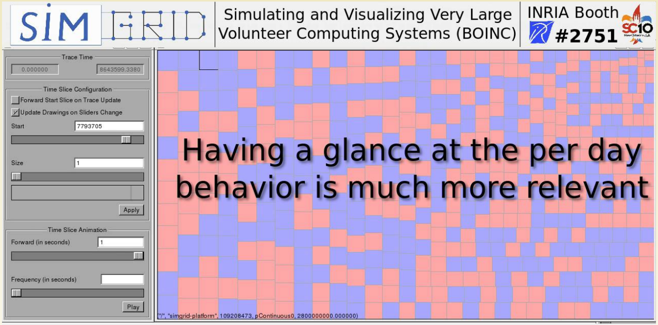
Treemap Examples



Treemap Examples



Treemap Examples

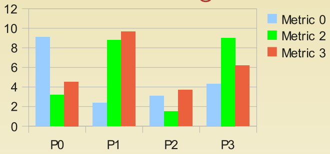


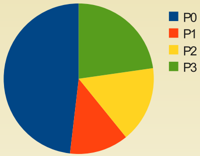
BOINC: 2 projects, 1000 volunteers, for 3 months

Outline

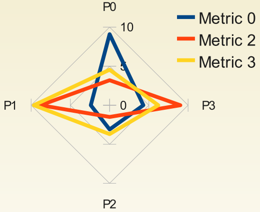
Classical Statistical Charts

Histograms and Pie-charts





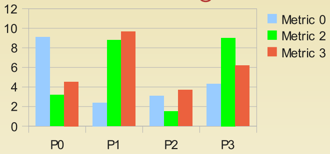
- Kiviats (a.k.a radar map). They consists of a sequence of equally distributed spokes, each one representing one of the monitored entities.

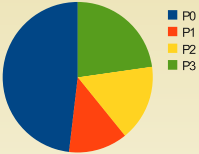


The spokes are connected by lines that limit an area whose shape can be used to detect imbalance.

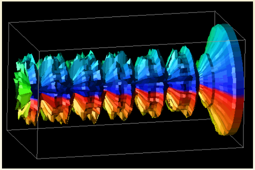
Such representations can have 3D versions as well (sexy! ;).

Histograms and Pie-charts





Kiviats (a.k.a radar map). They consists of a sequence of equally distributed spokes, each one representing one of the monitored entities.

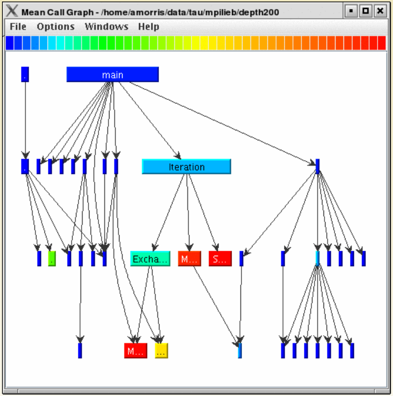


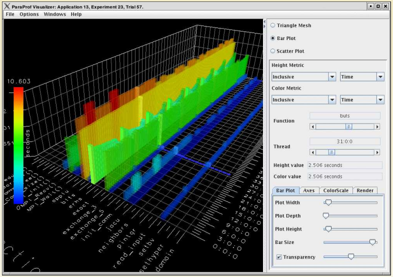
The spokes are connected by lines that limit an area whose shape can be used to detect imbalance.

Such representations can have 3D versions as well (sexy! ;).

Combining views

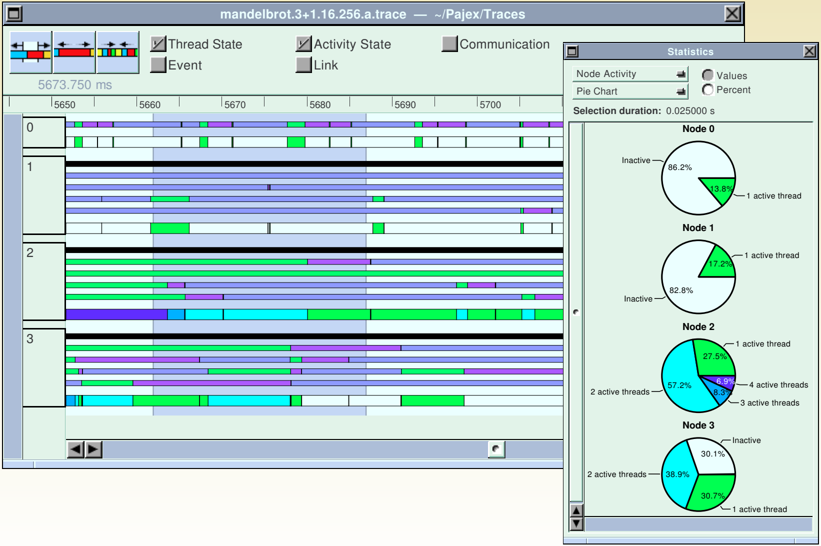
A good tool will be modular (truly) and enable you to combine views, to select entities, variables, . . .





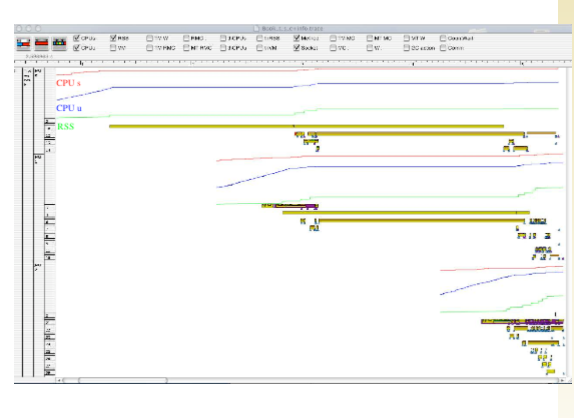
Combining views

A good tool will be modular (truly) and enable you to combine views, to select entities, variables, ...



Combining views

A good tool will be modular (truly) and enable you to combine views, to select entities, variables, . . .



Combining views

A good tool will be modular (truly) and enable you to combine views, to select entities, variables, . . .

The previous distributed Lagrangian video was also a good example.

Outline

Visual Analysis of I/O System Behavior for High End Computing

LSAP 2011

Figure 3: Color legend. The colors used in the timeline, scatter, and fisheye plots.

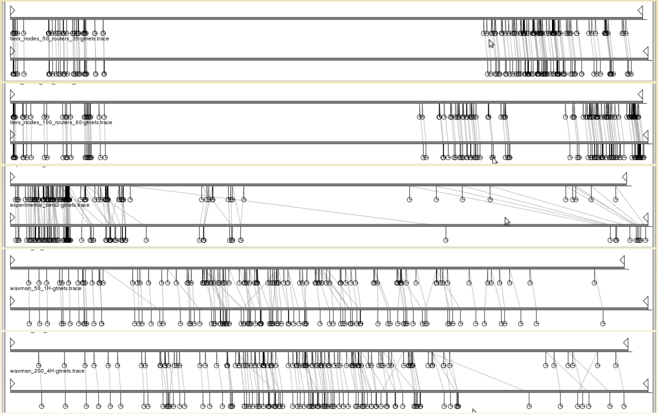
Figure 4: Point-based plot of I/O activity. The large point was added for illustrative purposes. The event it represents started at t_{start} and ended at t_{end}. The fact that this event was in execution at time $t_{current}$ is indicated by the point's location within the area between the logarithmic curve and asymptote corresponding to $t_{current}$.

Figure 6: Gantt chart with two fisheye zooms. A time range was selected, and this view uses sigmoid functions around the start and end times to expand the surrounding area

Visual Analysis of I/O System Behavior for High End Computing
LSAP 2011

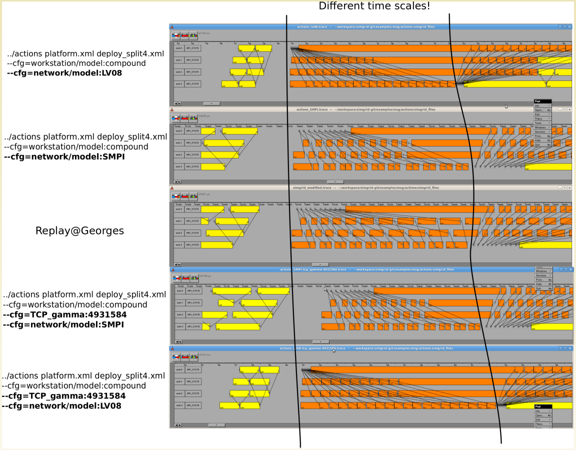
Outline

Trace Diff – Comparing SG with GTNetS



This is a diff from high-level events, which raises many time scale and synchronization issues.
From this, a finer diff could be made or maybe switching to another kind of view (e.g. a spatial view emphasizing the diff) ?

Trace Diff – Comparing two network models



This is a diff from high-level events, which raises many time scale and synchronization issues.
From this, a finer diff could be made or maybe switching to another kind of view (e.g. a spatial view emphasizing the diff) ?

Challenges

- Scalability: clusters, grids, clouds, P2P;
- Heterogeneous levels of traces (hierarchical organization)
- Seamless navigation at different scales
- Automatic pattern detection (e.g. late receive)
- Aggregation
- Trace comparison
- Synchronization issues in distributed systems
- Intrusiveness
- ...

Beware of visualization artifacts