Dynamic Network, Grid Computing and SCARIe

Damien Marchal

UNIVERSITEIT VAN AMSTERDAM

sara Computing & Networking Services

JIVE JOINT INSTITUTE FOR VLBI IN EUROPE

Context:

“How can we take profit of Starplane for the SCARIe project ?”

- how can we take profit of dynamic network in the context of grid ?
- how this can be applied to SCARIe ?

Dynamic Networks

My definition:

“A dynamic network is a network where link topology can be changed. Such networks offer to build virtual circuit with specific capability upon user request.”

Why dynamic networks arose:

- Network administrator want to do Traffic Engineering for a better best-effort routing;
- Some users need Quality of Service that best-effort cannot provide (TVoip);
- Some users need high capacity and QoS (eScience);

Dynamic Networks

My definition:

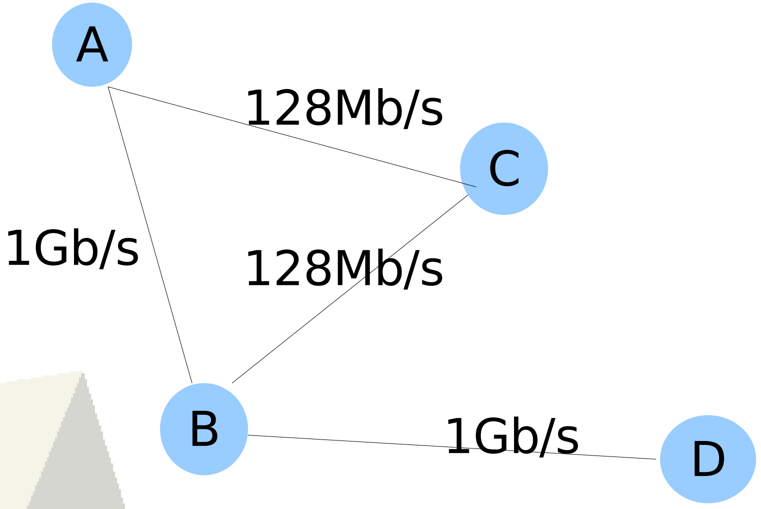
““A dynamic network is a network where link topology can be changed. Such network offer to build virtual circuit with specific capability upon user request.”

Examples of Virtual/Dynamic Network research initiative:

- GLIF:
  - Worldwide virtual organization to promote dynamic network based on optical multiplexing.
- Surfnet6 and Netherlight:
  - Netherland hybrid network offering allocation of Lightpath upon request.
- Starplane:
  - A DAS3 dynamic network offering allocation of Lightpath upon request.

Dynamic Networks

Dynamic network provides virtual circuit over physical netwo

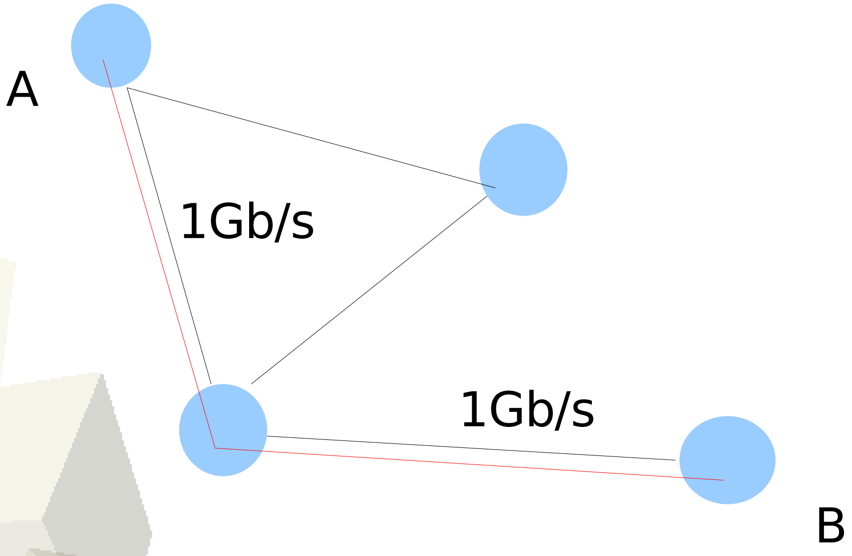


Physical network

Dynamic Networks

Dynamic network provides virtual circuit over physical netwo

Users can build their own virtual networks.

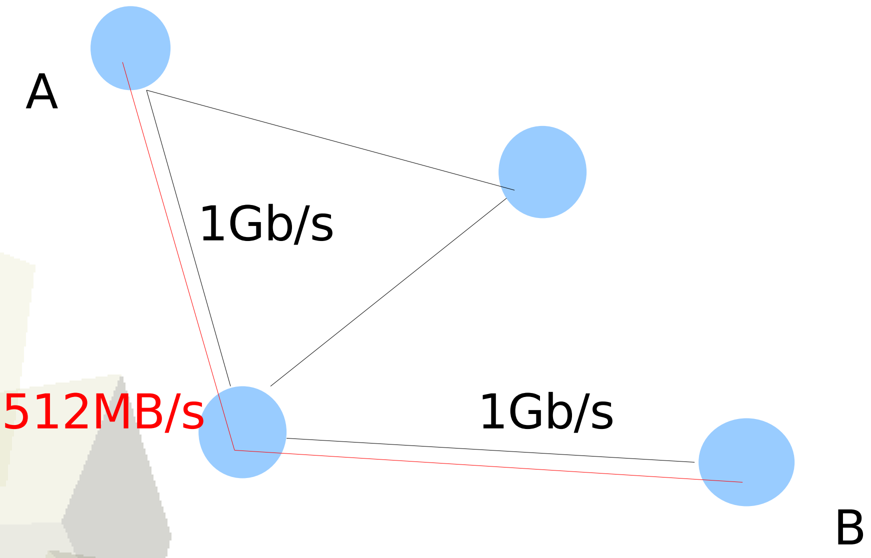


Physical network + virtual circuit between A-B

Dynamic Networks

Dynamic network provide virtual circuit over physical networ

Users can build their own virtual networks;
that match their Application Specific Requirement.



Physical network + virtual circuit + quality of servic

How user of a grid could take profit of dynamic network ?

To take profit of that ...

We need:

- infrastructure (they exist...see GLIF, surfnet6, Starplane);

- middleware and tools to build virtual circuit;
  at least a resource-allocator/scheduler aware of network

- middleware and tools that take profit of virtual circuit;
  dedicated version of GridFTP, ftp, scp over using virtual circuit t
  optimize data transmission.

- API that support dynamic networks;

- software methodologies to handle the increased complexity of dynamic network.

How user of a grid could take profit of dynamic network ?

Scientist that want to do big science...they need to use grid a dynamic network.

We need:

- infrastructure (they exist...see GLIF, surfnet6, Starplane);

- middleware and tools to build virtual circuit;
  at least a resource-allocator/scheduler aware of network

- middleware and tools that take profit of virtual circuit;
  dedicated version of GridFTP, ftp, scp over using virtual circuir t
  optimize data transmission.

- API that support dynamic networks;

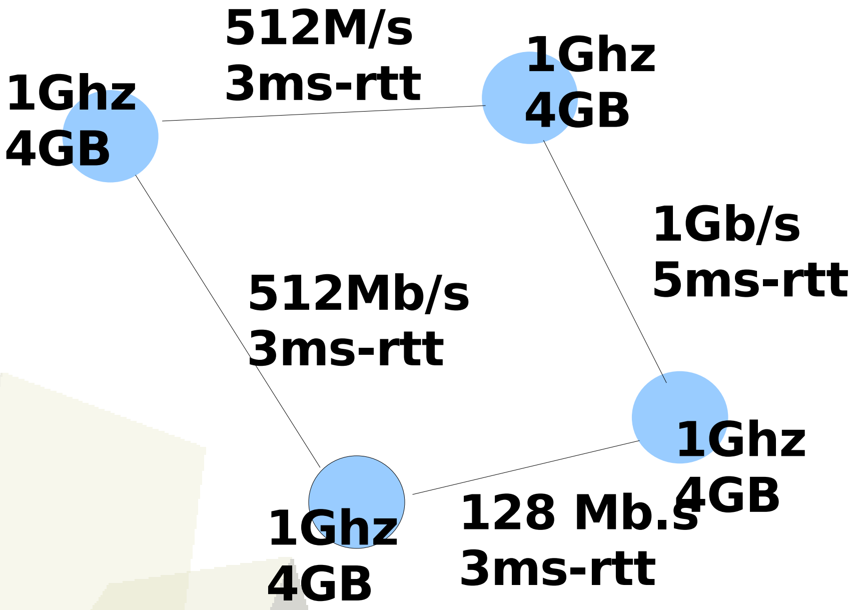
- software methodologies to handle the increased complexity of
  dynamic network.

How user of a grid could take profit of dynamic network ?

Simple model for network aware resource allocation...

Let's represent our resource by a graph:

- an node is a computation power or data storage element;
- an edge is a Virtual Path;

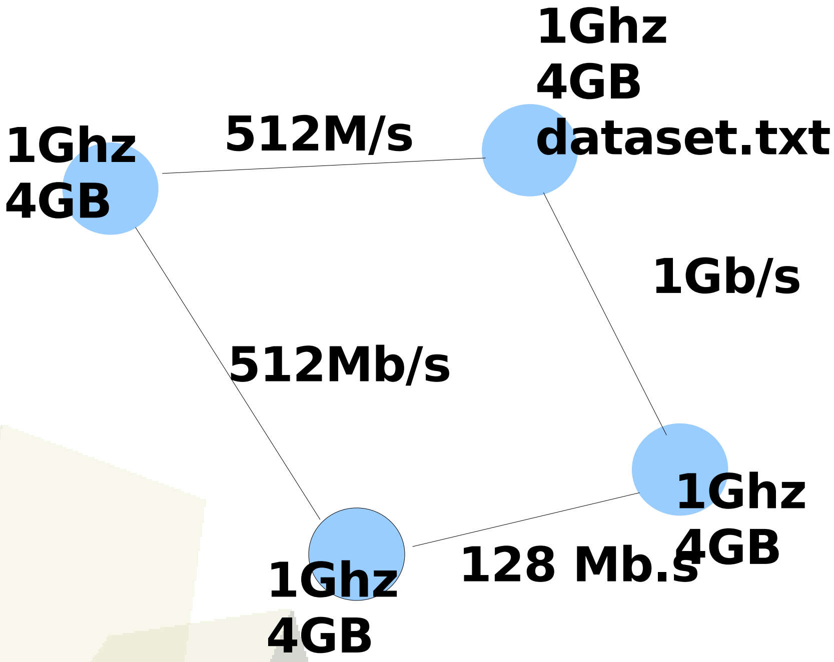


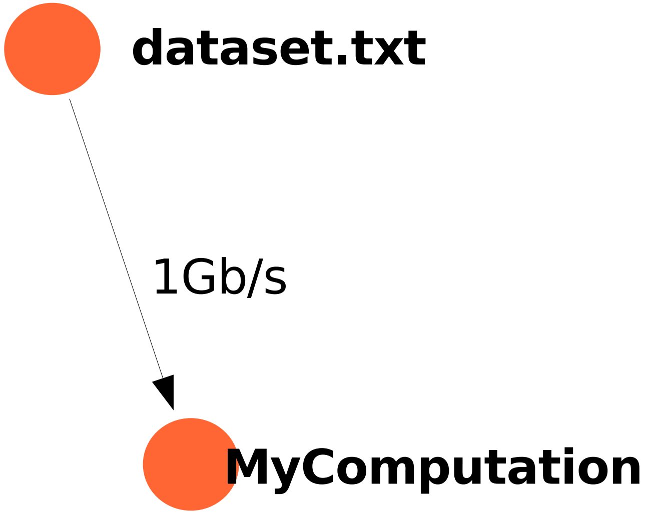
How user of a grid could take profit of dynamic network ?

Simple model for network aware resource allocation...

Let's represent our request by a graph:

- an node is a computation power or data storage el
- an edge is a Virtual Path;





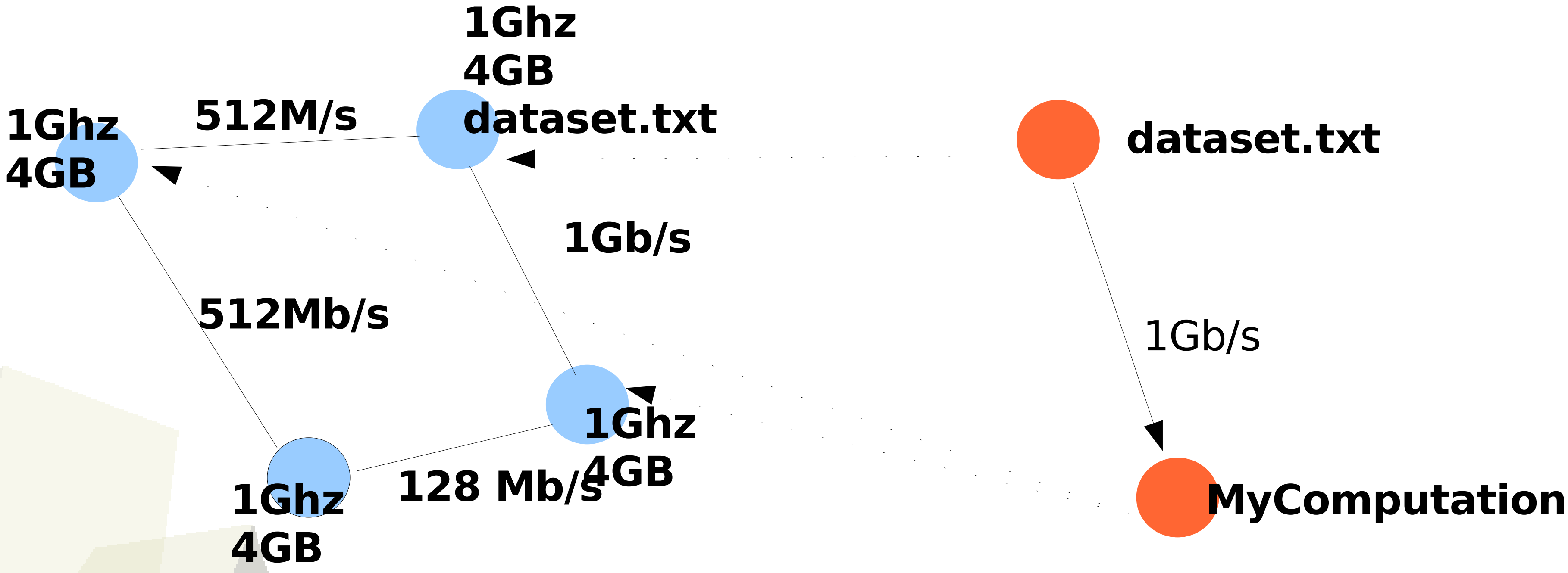
We need to map the request to the resources.

How user of a grid could take profit of dynamic network ?

Simple model for network aware resource allocation...

Let's represent our request by a graph:

- an node is a computation power or data storage el
- an edge is a Virtual Path;



We need to map the request to the resources.

This is a “subgraph isomorphism” test: At least NP-Complete

Don't expect an optimal solution

How user of a grid could take profit of dynamic network ?

Scientist that want to do big science...they need to use grid ar dynamic network.

We need:

- infrastructure (they exist...see GLIF, surfnet6, Starplane);

- middleware and tools to build virtual circuit;
  We need at least a resource-allocator/scheduler aware of netwo

- middleware and tools that take profit of virtual circuit;
  dedicated version of GridFTP, ftp, scp over using virtual circuir t optimize data transmission.

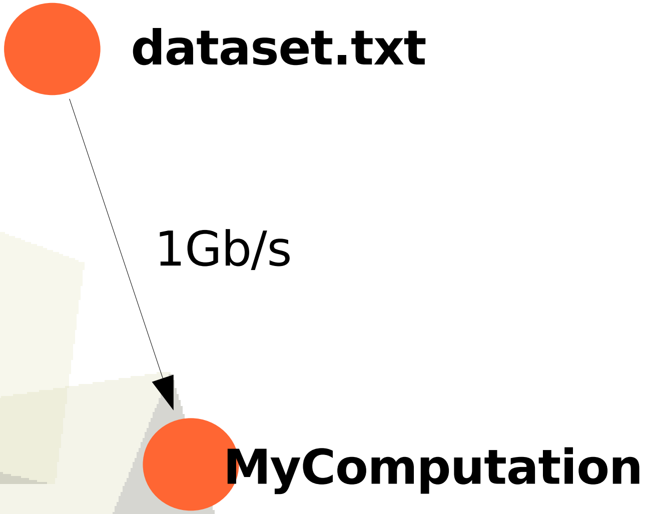
- API/Paradigm that support dynamic networks;

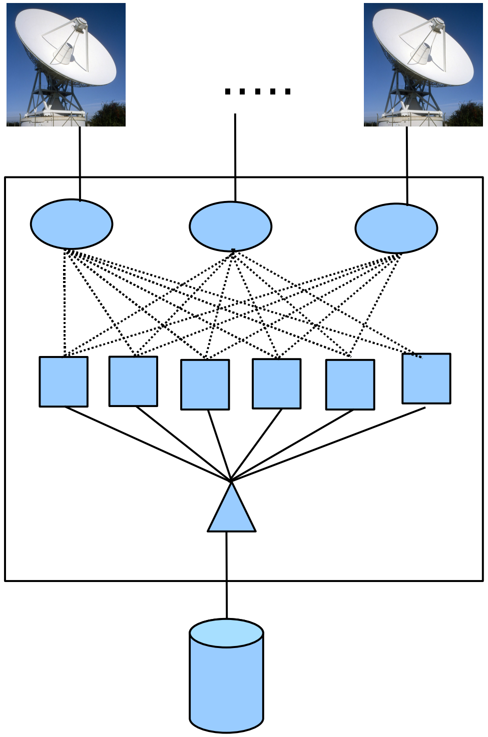
How user of a grid could take profit of dynamic network ?

Message Passing Library aware of Starplane

Message Passing Paradigm:

- the messages are explicitly exchanged between two computation elements;





How user of a grid could take profit of dynamic network ?

Message Passing Library aware of Starplane

Channels:

A Channel is like a MPI_Communicator + having the ability to be bounded
to a specific resource set (device, protocol, network).

when making a software... Identify a groups of channels :

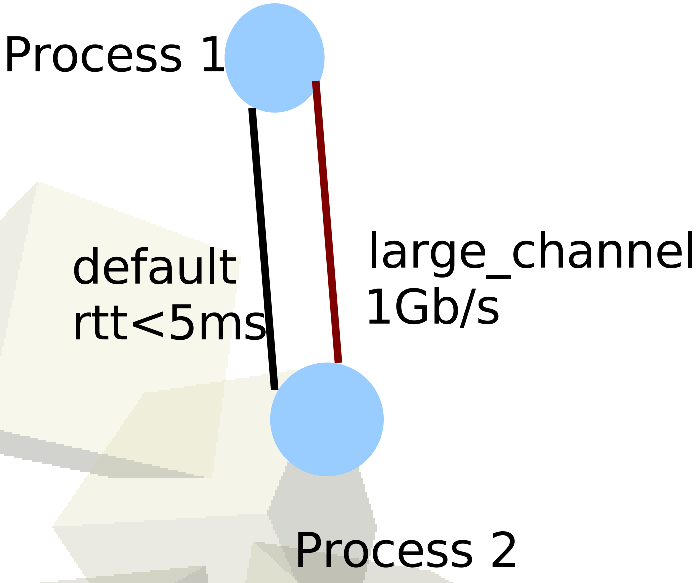
- same logical function (like MPI_Communicator)
- same requirement (bandwith,rtt)
- between same set of processes

How user of a grid could take profit of dynamic network ?

Message Passing Library aware of Starplane

Methodology:

1) Design your application using channels (here a scp like).



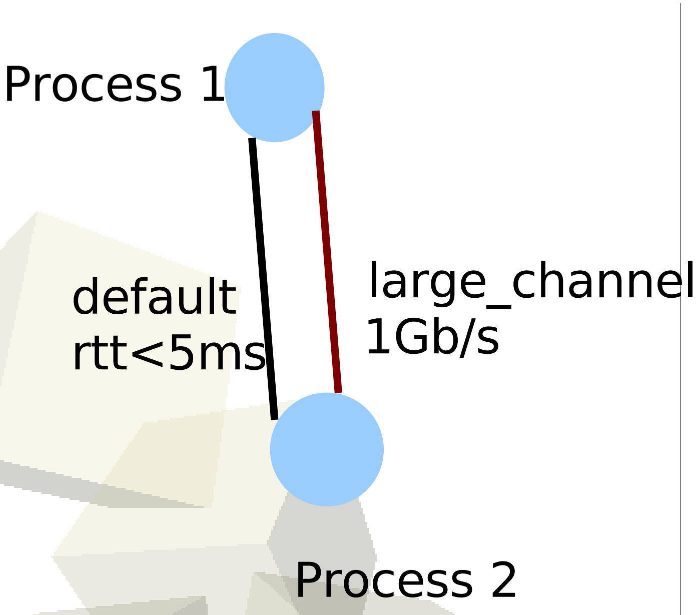
How user of a grid could take profit of dynamic network ?

Message Passing Library aware of Starplane

Methodology:

1) Design your application using channels;

2) The channel description is stored as an external file (the Application
Specific Topology).



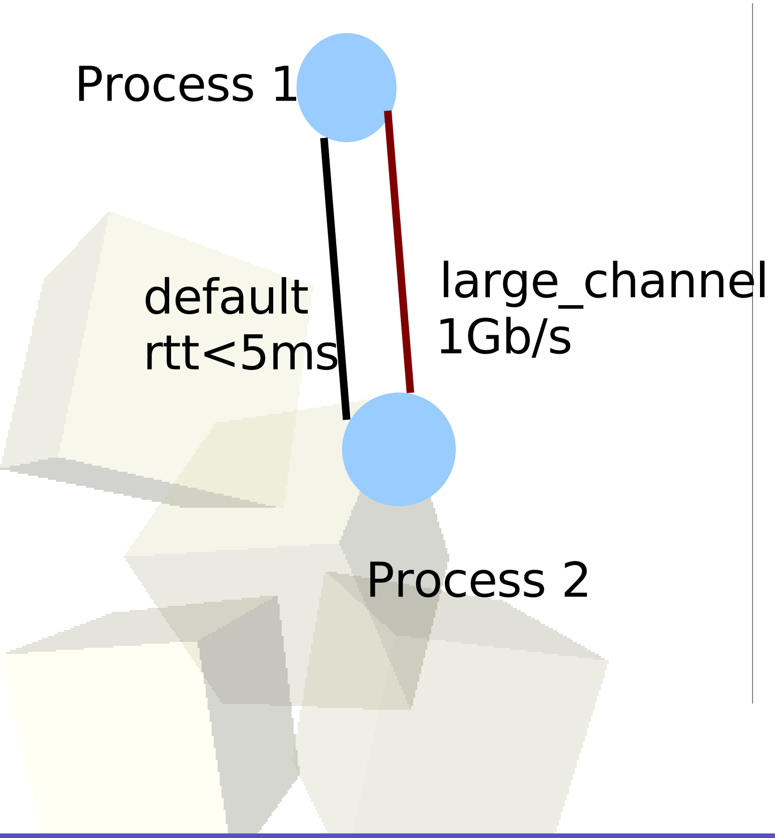
```
Process1:
Process2:
default_channel:
      rtt < 5ms
      members: Process1,
               Process2
large_channel:
      bandwith = 1Gb/s
      members: Process1,
               Process2
```

How user of a grid could take profit of dynamic network ?

Message Passing Library aware of Starplane

Methodology:

3) Mapping between channel and resource is done by resource allocator.



Process 1

default
rtt<5ms

large_channel
1Gb/s

Process 2

```
Process1:
Process2:
default_channel:
      rtt < 5ms
      members: Process1,
                     Process2
large_channel:
      bandwith = 1Gb/s
      members: Process1,
                     Process2
```

```
Process1 map to:
      Node234
Process2 map to:
       Node456

default channel map to:
      eth0, tcp
      Node234, Node456

large_channel map to:
      myrinet+starplane[red]
      Node 234, Node456
```

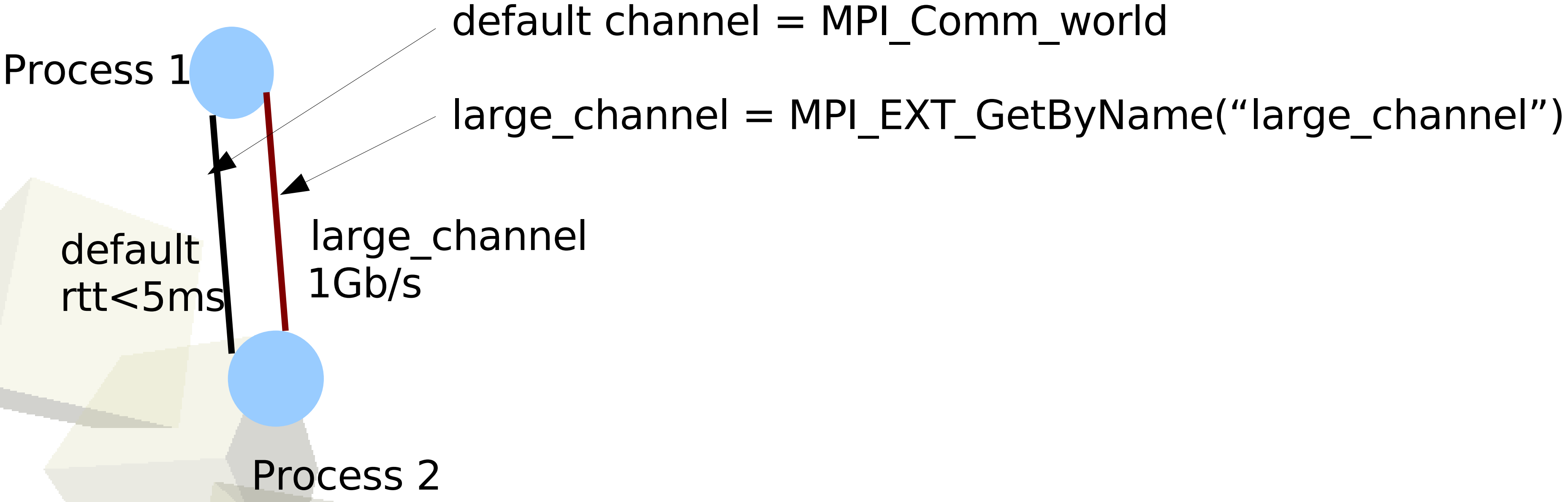
How user of a grid could take profit of dynamic network ?

Message Passing Library aware of Starplane

Methodology:

3) Mapping between channel and resource is done by resource allocator.

4) Application is started... the mapping is loaded to initialize the physical network connexions.


default channel = MPI_Comm_world
large_channel = MPI_EXT_GetByName(“large_channel”)
Process 1
default
rtt<5ms
large_channel
1Gb/s
Process 2

The application never know what is the underlaying tech

How user of a grid could take profit of dynamic network ?

Message Passing Library aware of Starplane

What we currently have:

- implementation of channel (with Madeleine library)
- implementation of channel over MPI (MPICH-Mad)
- some prototype of description language to request resources (
- some network resource description language (NDL is very den

What we are missing:

- a resource allocator/scheduler that map the AST to subset of N
- grid aware set of tools.

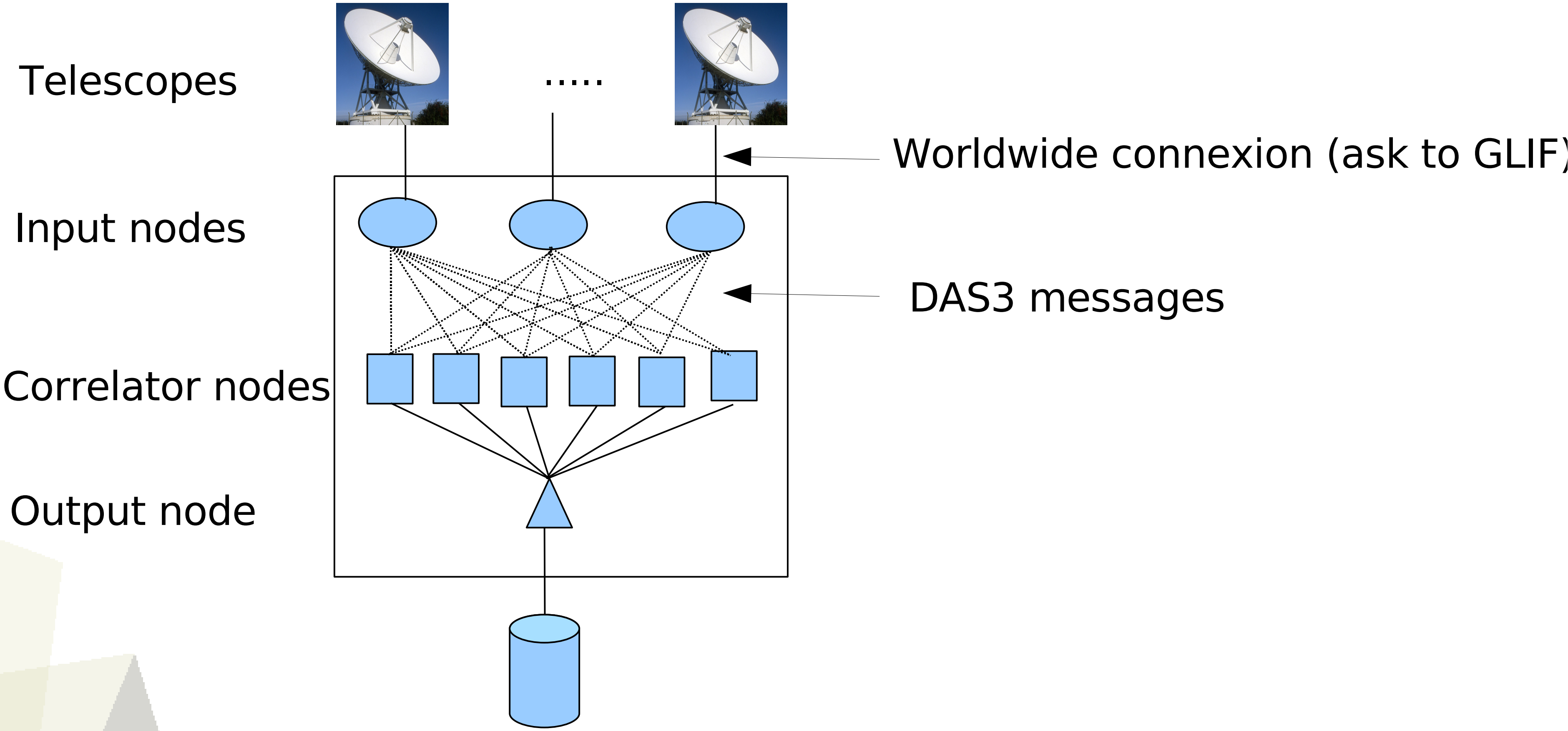
Context:

“How can we take profit of Starplane for the SCARIe project ?”

- how can we take profit of dynamic network in the context of grid ?
- **how this can be applied to SCARIe ?**

SCARIe

An exemple of usage of dynamic network for static application



Telescopes

.....

Worldwide connexion (ask to GLIF)

Input nodes

DAS3 messages

Correlator nodes

Output node

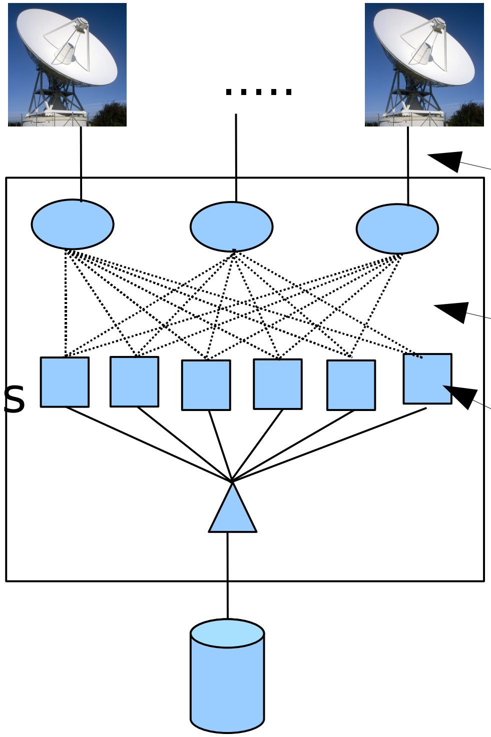
What is drawn in this picture is a graph with:

- Node that represents computation process

- Edges that represents streams of data between the computation process

SCARIe

An exemple of usage of dynamic network for static application



Telescopes

Input nodes

Correlator nodes

Output node

To equal the hardware correlator we need:

16 streams of 1Gbps

16 * 1Gbps of data

2 Tflops of computation power

By knowing the number of antenna, the number of channels it is relatively easy to estimate for all Edges and Nodes of this graph the amount of resources needed.

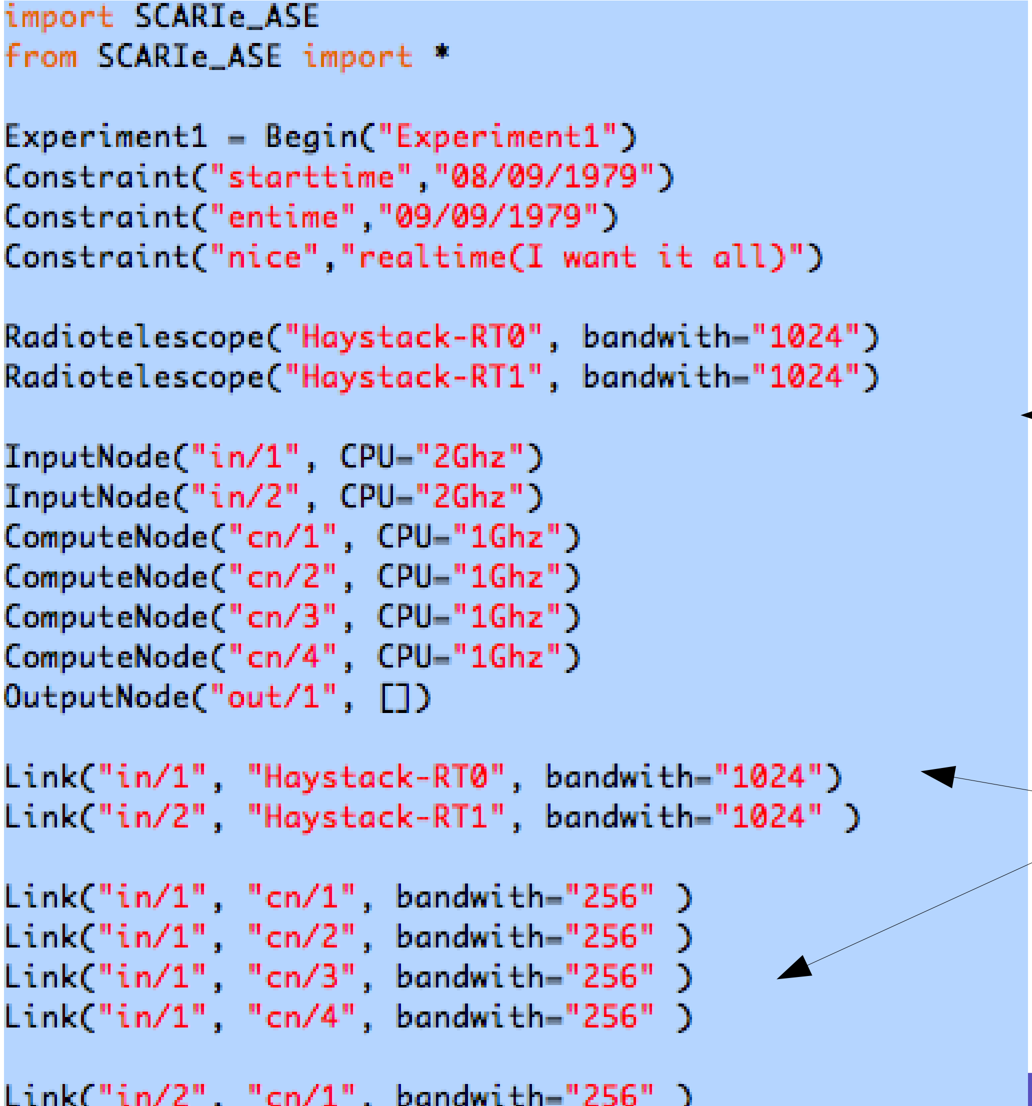
The application has *STATIC Requirement*

SCARIe

An exemple of usage of dynamic network for static application

First version of SCARIE on top of Starplane:

1) we provide a description of the application topology;



```
import SCARIe_ASE
from SCARIe_ASE import *

Experiment1 = Begin("Experiment1")
Constraint("starttime","08/09/1979")
Constraint("entime","09/09/1979")
Constraint("nice","realtime(I want it all)")

Radiotelescope("Haystack-RT0", bandwith="1024")
Radiotelescope("Haystack-RT1", bandwith="1024")

InputNode("in/1", CPU="2Ghz")
InputNode("in/2", CPU="2Ghz")
ComputeNode("cn/1", CPU="1Ghz")
ComputeNode("cn/2", CPU="1Ghz")
ComputeNode("cn/3", CPU="1Ghz")
ComputeNode("cn/4", CPU="1Ghz")
OutputNode("out/1", [])

Link("in/1", "Haystack-RT0", bandwith="1024")
Link("in/2", "Haystack-RT1", bandwith="1024" )

Link("in/1", "cn/1", bandwith="256" )
Link("in/1", "cn/2", bandwith="256" )
Link("in/1", "cn/3", bandwith="256" )
Link("in/1", "cn/4", bandwith="256" )

Link("in/2", "cn/1", bandwith="256" )
```

Example of AST

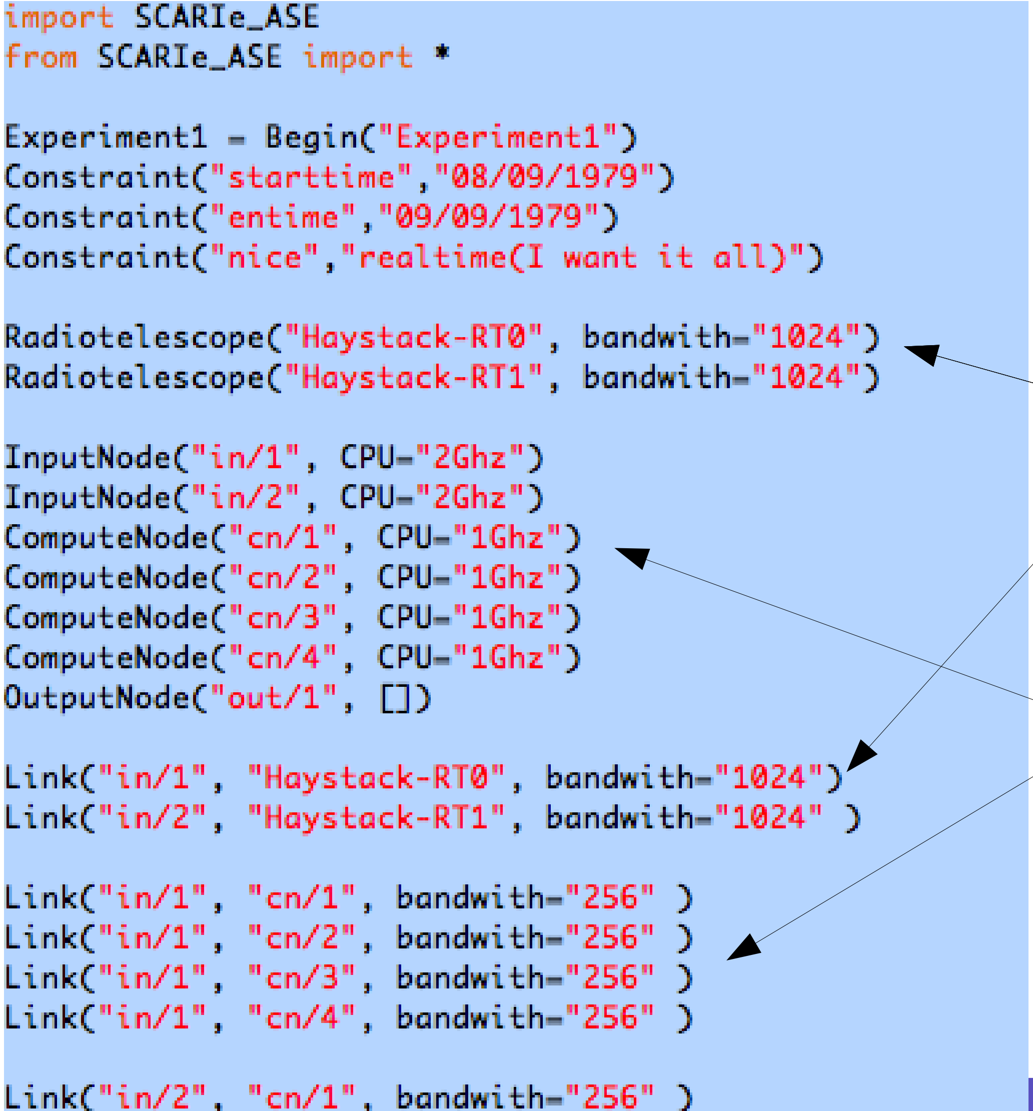
List of point2point channels

SCARIe

An exemple of usage of dynamic network for static application

First version of SCARIE on top of Starplane:

1) we provide a description of the application topology;



```
import SCARIe_ASE
from SCARIe_ASE import *

Experiment1 = Begin("Experiment1")
Constraint("starttime","08/09/1979")
Constraint("entime","09/09/1979")
Constraint("nice","realtime(I want it all)")

Radiotelescope("Haystack-RT0", bandwith="1024")
Radiotelescope("Haystack-RT1", bandwith="1024")

InputNode("in/1", CPU="2Ghz")
InputNode("in/2", CPU="2Ghz")
ComputeNode("cn/1", CPU="1Ghz")
ComputeNode("cn/2", CPU="1Ghz")
ComputeNode("cn/3", CPU="1Ghz")
ComputeNode("cn/4", CPU="1Ghz")
OutputNode("out/1", [])

Link("in/1", "Haystack-RT0", bandwith="1024")
Link("in/2", "Haystack-RT1", bandwith="1024" )

Link("in/1", "cn/1", bandwith="256" )
Link("in/1", "cn/2", bandwith="256" )
Link("in/1", "cn/3", bandwith="256" )
Link("in/1", "cn/4", bandwith="256" )

Link("in/2", "cn/1", bandwith="256" )
```

Element that shoud be sticked to specific resource.

Location-free elements

SCARIe

An exemple of usage of dynamic network for static application

First version of SCARIE on top of Starplane:

1) Channels are expressed, we have a topological description of the application (AST).

2) The AST is submitted to the resource scheduler...

3) a *Resource Mapping* file is returned containing:
   - for each computation process a computing node.
   - for each channel in the request the network path that is supposed to be used.

4) When the application is started:
we are sure that the resources are available;
we are sure that the application fit into the resource required.
....this is good for real-time application...

SCARIe

An exemple of usage of dynamic network for static application

Second version of SCARIE on top of Starplane:
how to make a a better demonstration of the dynamic capabilities of Starplane.

The idea... SFXC use time slicing to distribute job.

At cluster level:
- timeslice is used to distribute computation

At grid level:
- chunk of timeslices.

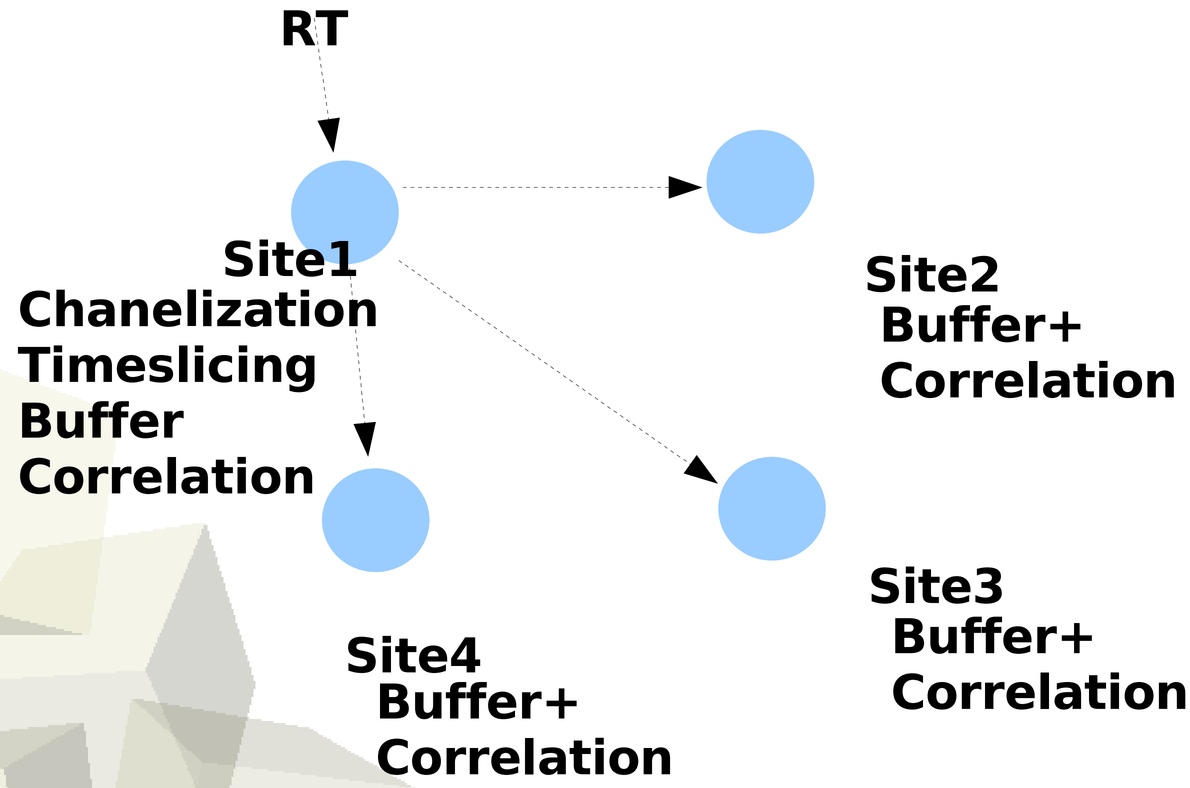
How can we distribute timeslices sequentially over the different cluster sites ?

SCARIe

An exemple of usage of dynamic network for static application

Second senario of SCARIE on top of Starplane:

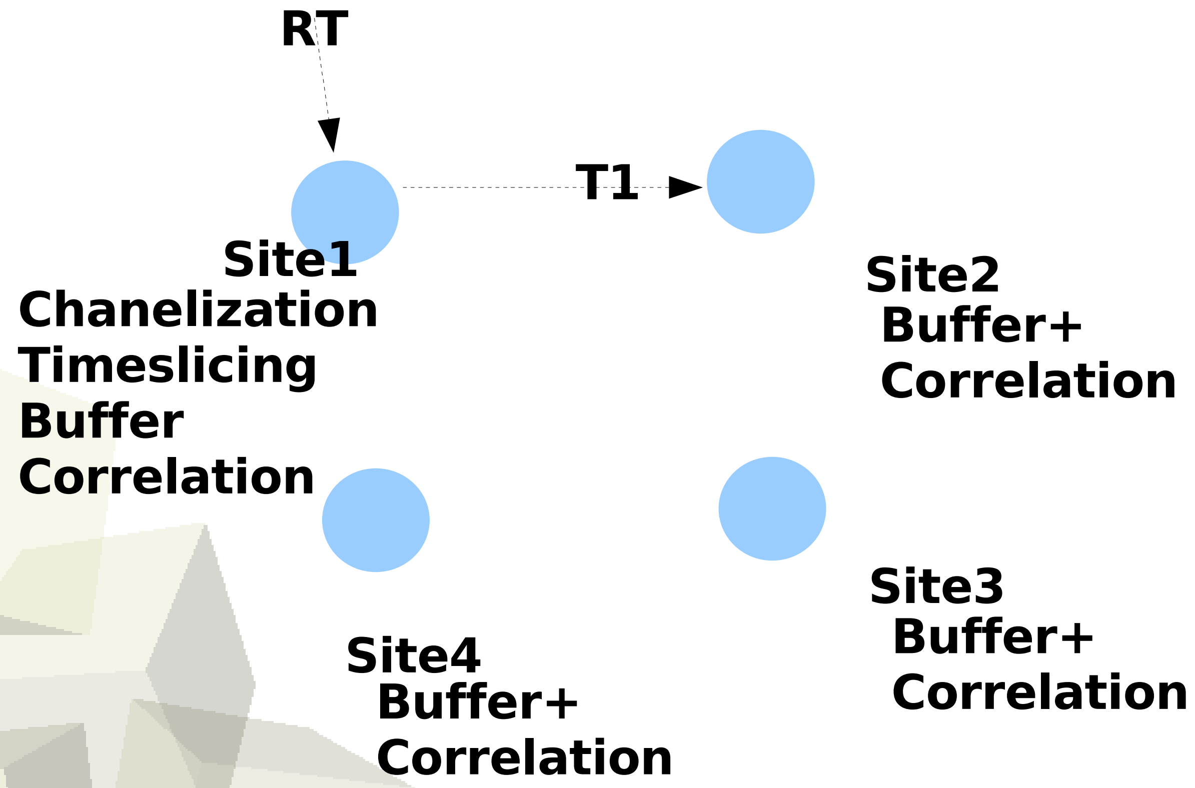
An attempt to a better demonstration of the dynamic capabilities of Starplane.



SCARIe

An exemple of usage of dynamic network for static application

Second senario of SCARIE on top of Starplane:



RT

T1

Site1
Chanelization
Timeslicing
Buffer
Correlation

Site2
Buffer+
Correlation

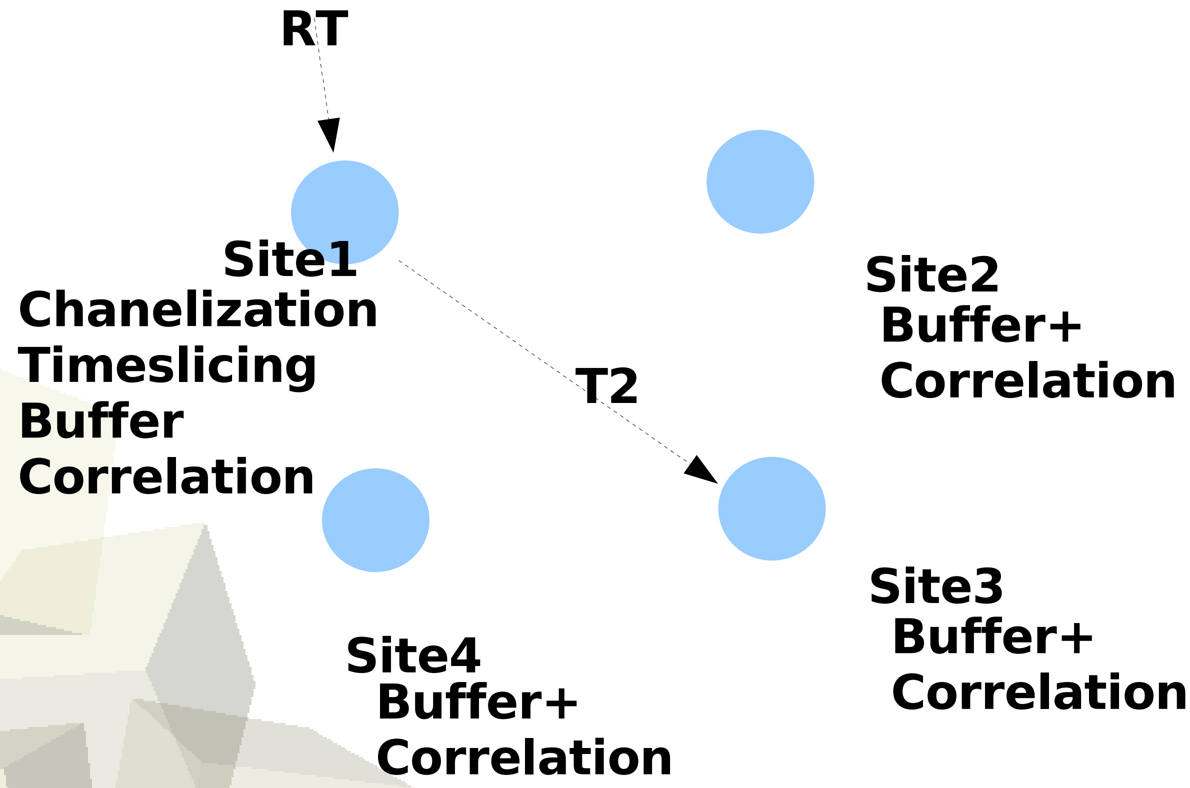
Site3
Buffer+
Correlation

Site4
Buffer+
Correlation

SCARIe

An exemple of usage of dynamic network for static application

Second senario of SCARIE on top of Starplane:



RT

Site1
Chanelization
Timeslicing
Buffer
Correlation

T2

Site2
Buffer+
Correlation

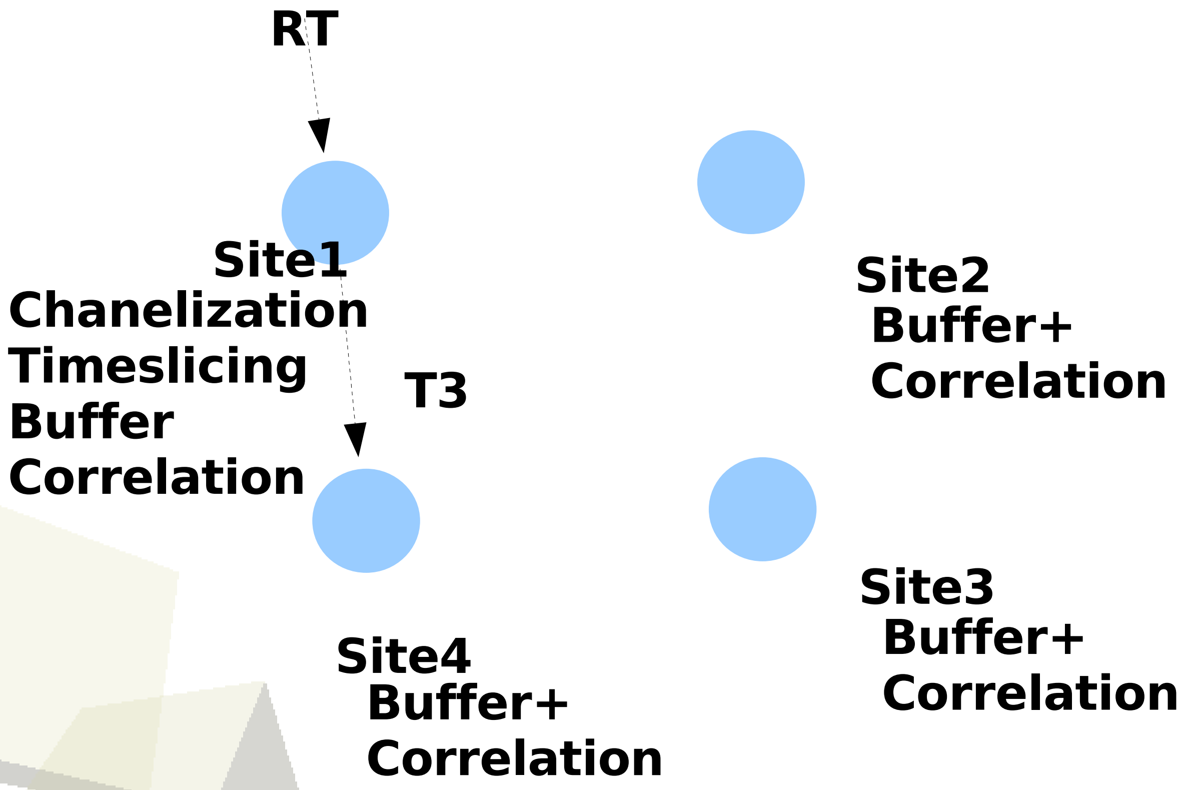
Site3
Buffer+
Correlation

Site4
Buffer+
Correlation

SCARIe

An exemple of usage of dynamic network for static application

Second senario of SCARIE on top of Starplane:



RT

Site1
Chanelization
Timeslicing
Buffer
Correlation

T3

Site2
Buffer+
Correlation

Site3
Buffer+
Correlation

Site4
Buffer+
Correlation

SCARIe

The SFXC software: static point of view

Second senario of SCARIE on top of Starplane:

This demonstrate the dynamically controlled lighpath of Sta

Fast light-path modification allow to keep the size of the bu small (less than the main memory of the nodes: 4GB);

Inter-site communication is only streaming, no message dia

Load distribution can be adapted by selecting different size timeslices;

Finished

That is all for me...

...

...

....

What about a demo of this for SC07 ?

....

....

....

....

What do we want to demonstrate ?

....

....

....

....