



Mixing Cloud and Grid Resources for Many Task Computing

David Abramson

Monash e-Science and Grid Engineering Lab (MeSsAGE Lab)
Faculty of Information Technology

Science Director: Monash e-Research Centre

ARC Professorial Fellow

Introduction



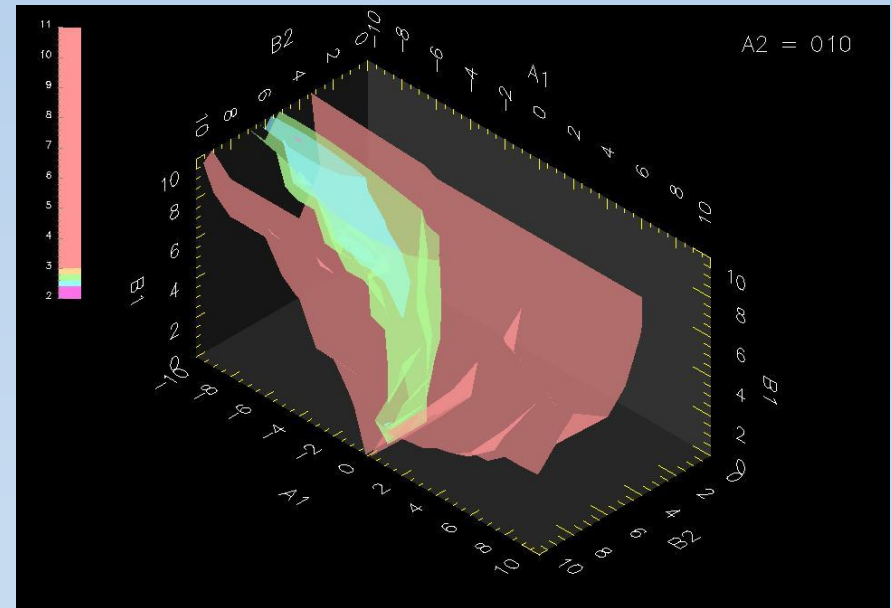
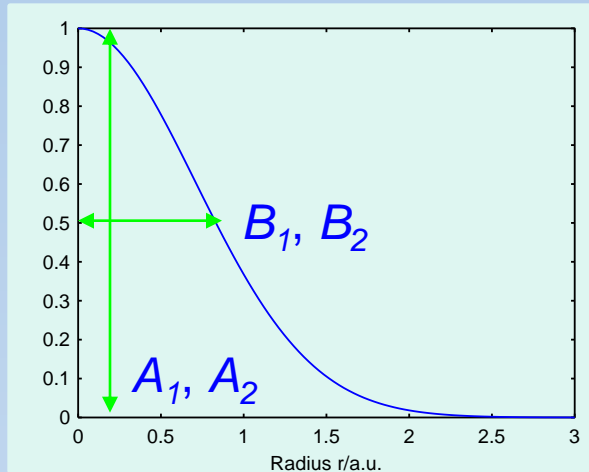
- A typical MTC Driving Application
- The Nimrod tool family
- Things the Grid ignored
 - Deployment
 - Deadlines (QoS)
- Clusters & Grids & Clouds
- Conclusions and future directions



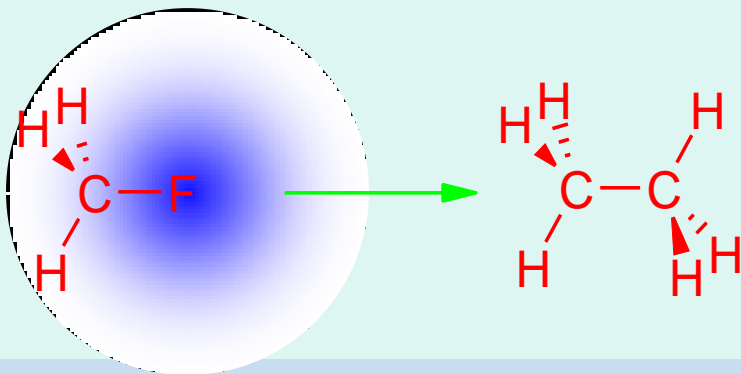
A Typical MTC Driving Application

A little quantum chemistry

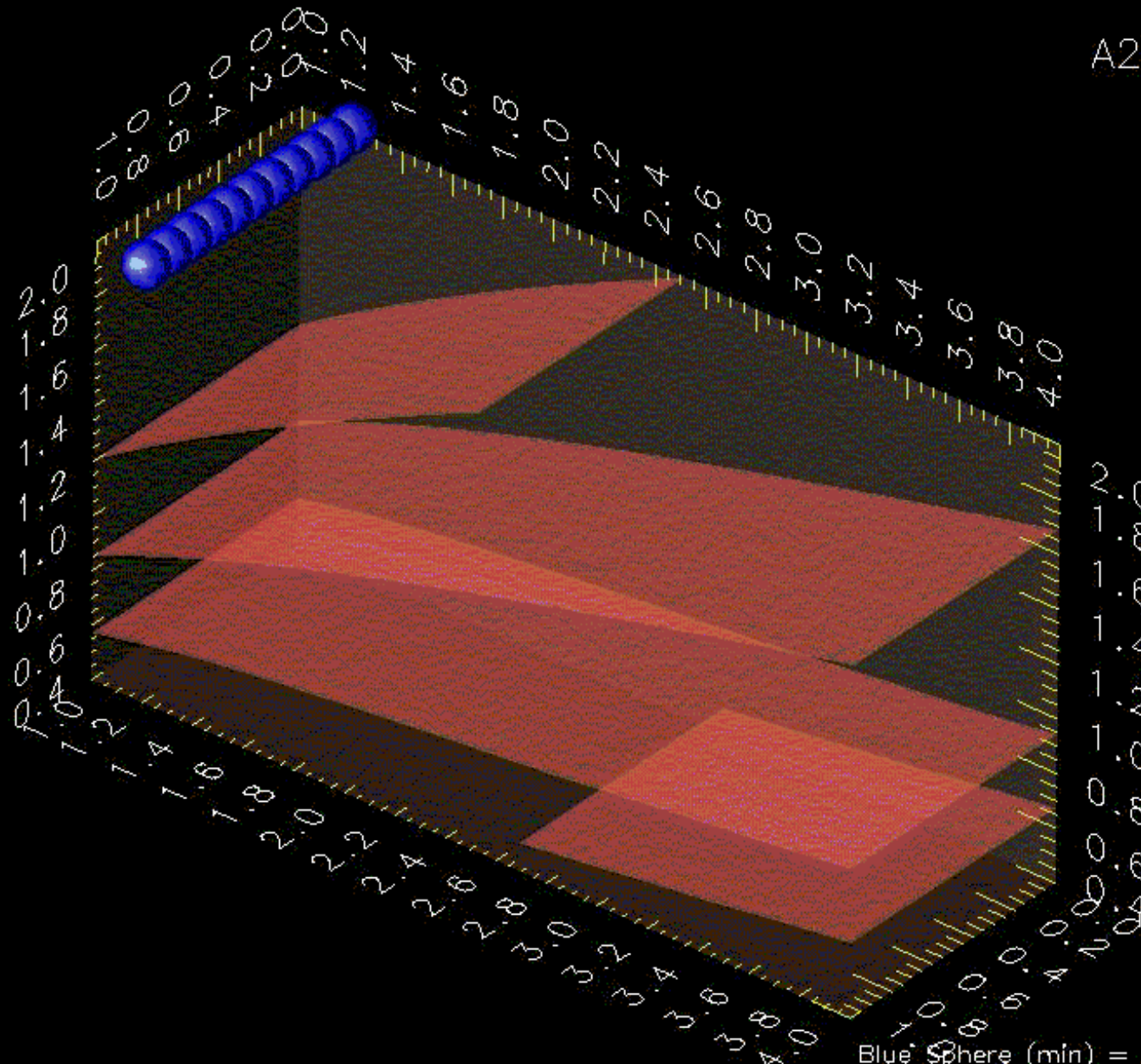
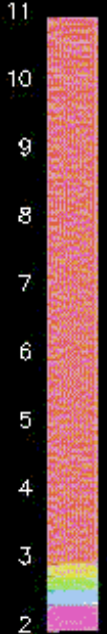
Wibke Sudholt, Univ Zurich



$$U_{\text{eff}}(r) = A_1 \exp(-B_1 r^2) + A_2 \exp(-B_2 r^2)$$

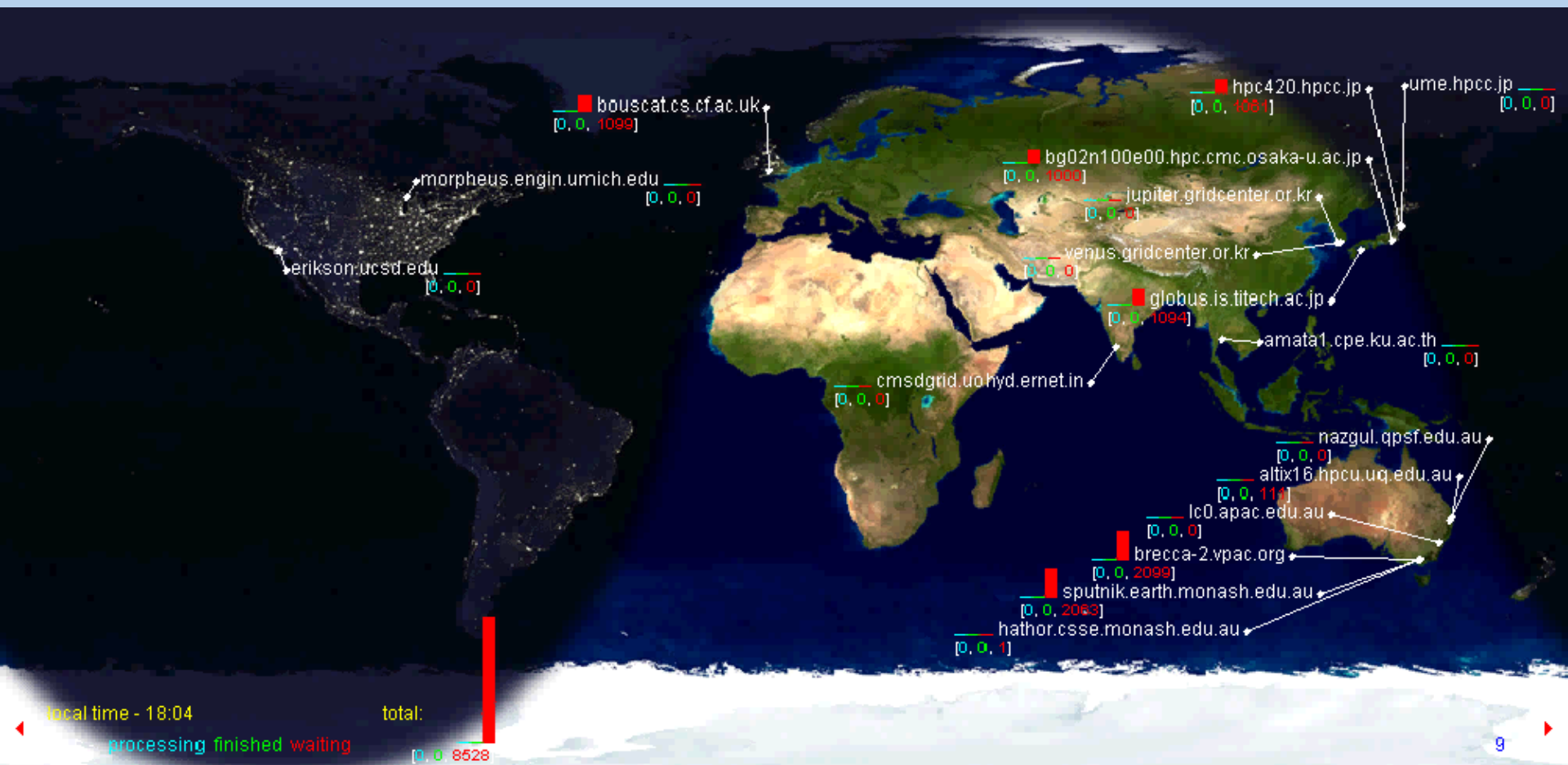


A2 = 00.0



Blue Sphere (min) = 2.671456

SC03 testbed



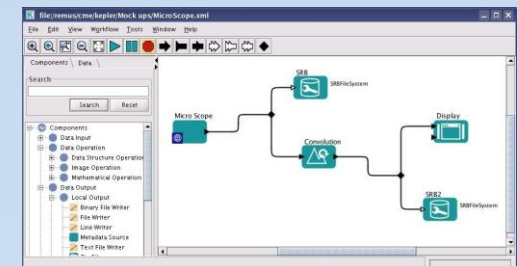
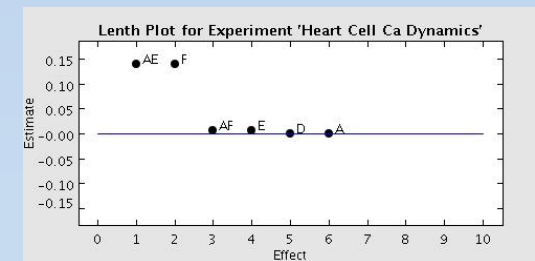
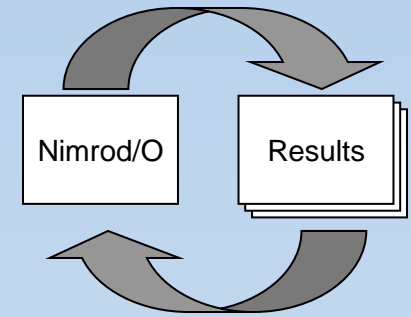
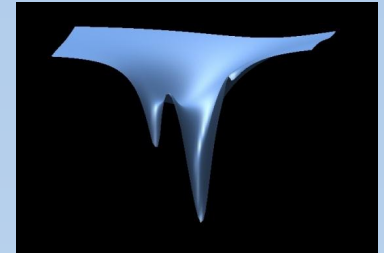


The Nimrod Tools Family

Nimrod supporting "real" science



- A full parameter sweep is the cross product of all the parameters (Nimrod/G)
- An optimization run minimizes some output metric and returns parameter combinations that do this (Nimrod/O)
- Design of experiments limits number of combinations (Nimrod/E)
- Workflows (Nimrod/K)



Legacy Nimrod family



Plan File

Nimrod Portal

Nimrod/O

Nimrod/E

Nimrod/G

Actuators

Grid Middleware

Nimrod Portal

Parameter Section

<input type="checkbox"/>	<input type="checkbox"/>	temp	label	<input type="text"/>	float	range	from	<input type="text" value="200"/>	to	<input type="text" value="300"/>	Points	<input type="text" value="2"/>
<input type="checkbox"/>	<input type="checkbox"/>	pressure	label	<input type="text"/>	float	range	from	<input type="text" value="5000"/>	to	<input type="text" value="6000"/>	Points	<input type="text" value="4"/>
<input type="checkbox"/>	<input type="checkbox"/>	concent	label	<input type="text"/>	float	range	from	<input type="text" value="0.002"/>	to	<input type="text" value="0.005"/>	Points	<input type="text" value="2"/>
<input type="checkbox"/>	<input type="checkbox"/>	material	label	<input type="text"/>	text	select	<input type="text"/>	<input type="button" value="Add"/>	<input type="text" value="Fe"/>	<input type="button" value="Remove"/>		

Add a new parameter

Add a comment

Tasks Section

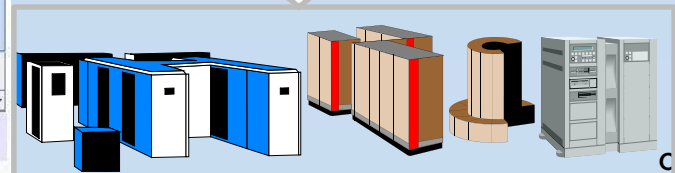
- task rootstart
- task nodestart
- task main
- task nodefinish
- task rootfinish

Add a new task

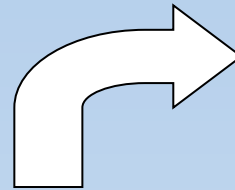
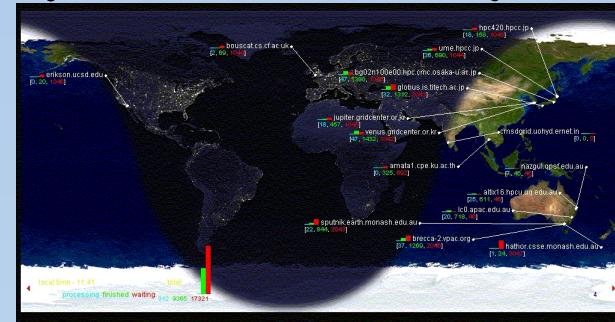
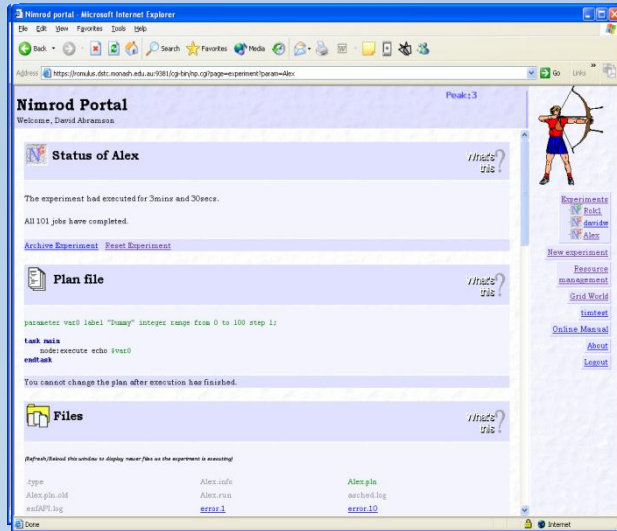
Save

Cancel and Reload

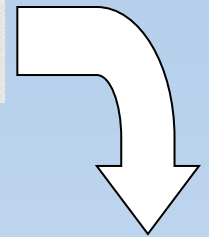
Text mode



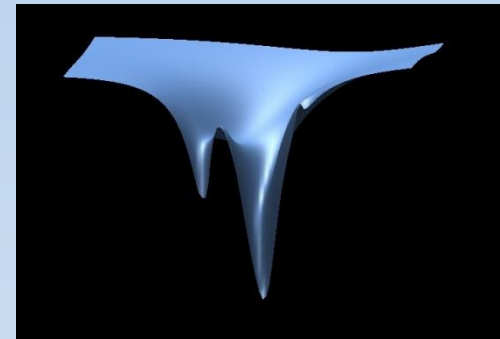
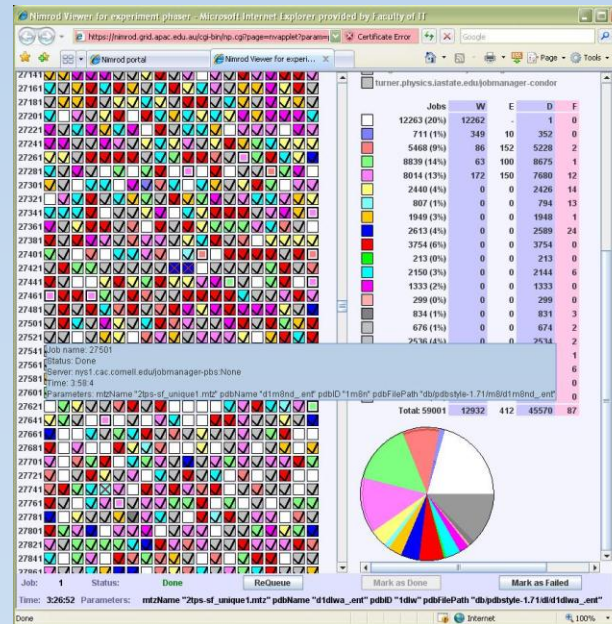
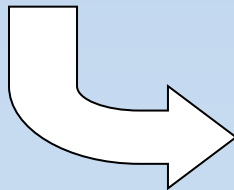
Nimrod Development Cycle



Sent to available machines



Prepare Jobs using Portal



Results displayed & interpreted

Jobs Scheduled Executed Dynamically

```
root@east-
Terminal Tabs Help
8 0/ 0 18/ 18
0 1/ 8 1/ 8
3 0/ 0 53/ 53
[screen 8: test] blair@nimrod1:~
Terminal Tabs Help
checked command 30 ready:False
checked command 70 ready:False
:1
e internal queue:
ing:0
81707500 pausing for 17.2020218372
main thread
:1
main thread
e internal queue:
ing:0
parallel:total
81707500 pausing for 18.2319529057
main thread
checked command 72 ready:True
main thread
unchecked command 72 ident:1236
checked command 70 ready:True
main thread
unchecked command 70 ident:1241
checked command 30 ready:True
tj$ ssh-keyscan -t rsa ec2-174-129-155-
204.compute-1.amazonaws.com SSH-2.0-Ope
4.compute-1.amazonaws.com ssh-rsa AAAAB
xQD75xcN6nRoedkR0TnJSLMurlmrCrvvoTA17kq
fCm32YJmFTX0Aaamtaq1lIAw//SnaEz0cFIIGfC
emljXRqo40VEIhdWf9i9YAt1G7ikgRe8LJY2MjQ
9oKQ==
tj$
tj$
tj$ tailf ~/nimrod/ec2.log
unchecked command 31 ident:1251
:1
e internal queue:
ing:0
:1
e internal queue:
ing:0
81707500 pausing for 18.2695720196
main thread
main thread
main thread
:1
e internal queue:
ing:0
81707500 pausing for 18.2067079544
main thread
main thread
:1
urce 3081761388
e int@fnaT`Q080e:
ing:0
81707500 pausing for 18.2608129978
main thread
```

Nimrod portal - Swiftfox

File Edit View History Bookmarks Tools Help

https://messagelab.monash.edu.au/NimrodARCS/cgi-bin/np-nl.cgi?page=experiments?window=

Gmail Latest BBC Headl... Zazz Cricket Aus Chaser foobar2000 Inbox - blair.bethwai... Most Visited Latest Headlines

Disable Cookies CSS Forms Images Information Miscellaneous Outline Resize Tools View Source Options

Inbox - ... ITS Mes... Nimr... pyar.py... Build a ... Amazo... Cloud E... A - Z El... http...ges http...0& boto -P... http...txt


Boto Amazo... The Pyt... 11.1. os... Amazo... Amazo... 36.16. c... 18.5. po... 18.1. su... /trunkjni... Eucalyp... Eucalyp...

17.2. th... Amazo... MeSsaA... 7. Built-i... Testing ... ec2 bot... Amazo... Tip: Get...

Nimrod Portal

Welcome Blair Bethwaite (you are an administrator)

Peak:8



Status of test

What's this?

The experiment has executed for 6mins and 0secs.

The experiment has started. 29 jobs have been completed, there are 23 jobs waiting, 38 jobs executing and 0 jobs failed.

[Check for grid errors](#)

[Pause Experiment](#)

Plan file

What's this?

```
parameter x integer range from 7 to -1 step 1;
parameter y float range from 0.125 to 1.25 step 0.125;

task main
  copy work.py node:.
  node:execute /bin/chmod +x work.py
  node:execute /bin/hostname >> output
  node:execute /bin/date >> output
  node:execute /bin/pwd >> output
  node:execute /bin/echo "Working..." >> output
  node:execute ./work.py ${x} ${y} >> output
  copy node:output output.${jobname}.x${x}.y${y}
endtask
```

You cannot change the plan after execution has started.

File Machine Help
New Settings Show Discard

- Hardy Saved
- Intrepid Powered Off
- NG Dev Powered Off
- XPL (Pre SFU) Running**

Details

Gene
Name
OS Ty
Base
Video
Boot C
ACPI:
IO AP
VT-x/A
Neste
PAE/N
3D Ac

Hard
IDE Pr

CD/D
Image

Flopp
Not m

Audic
Host C
Contr

Nimrod Viewer for experiment test - Swiftfox

https://messagelab.monash.edu.au/NimrodARCS/cgi-bin/np-ml.cgi?page=nvapplet?param=test?window=main

1
21
41
61
81

none
 nimrodtest.q@https://east-globus.enterprisegrid.edu.au:8443/SGE@
 cloud1//home1/blair/ec2/david-id/home1/blair/ec2/david-secret2:5:m1.small:kami-0d729464

Jobs	W	E	D
0 (0%)	0	-	0
67 (74%)	7	32	28
23 (25%)	5	13	5
Total: 90	12	45	33

Job: 1 Status: Executing ReQue... Mark as Do... Mark as Fail...

Time: 0:00:00 Parameters: x "7" y "0.125"

Applet nimrodviewer.Applet started

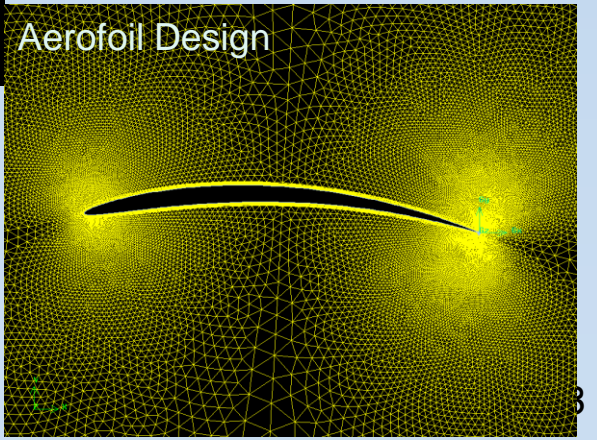
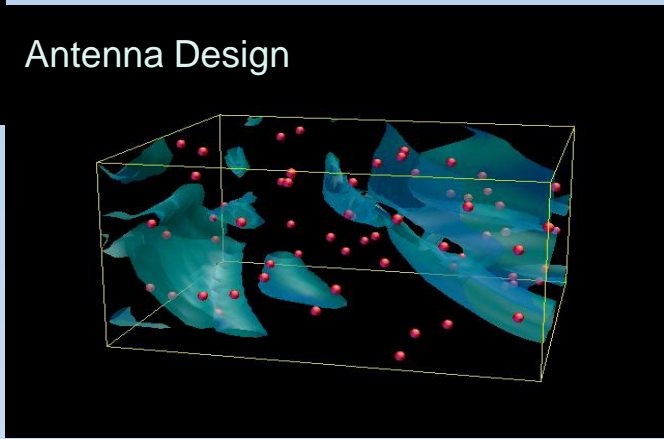
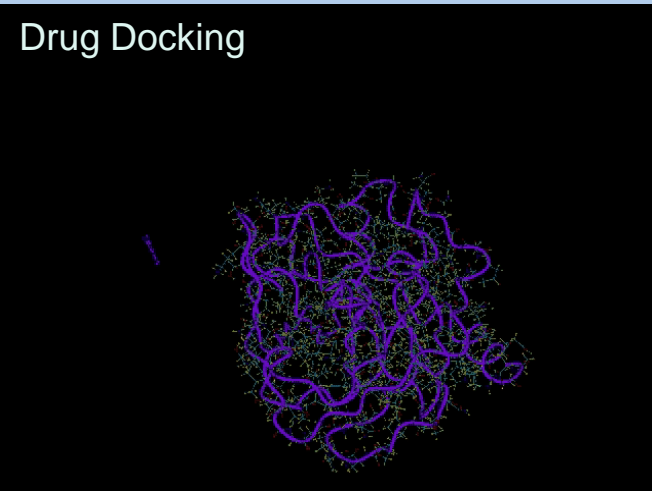
messagelab.monash.edu.au Open Notebook zotero

- dmesg WaspFactory.txt
- gt4-test.sh EVOPlayer.jnlp
- pcm9_linux e169g-switch_0.3 all.deb
- linux-phc-0.3.0-pre1 temp
- pslinuxv180

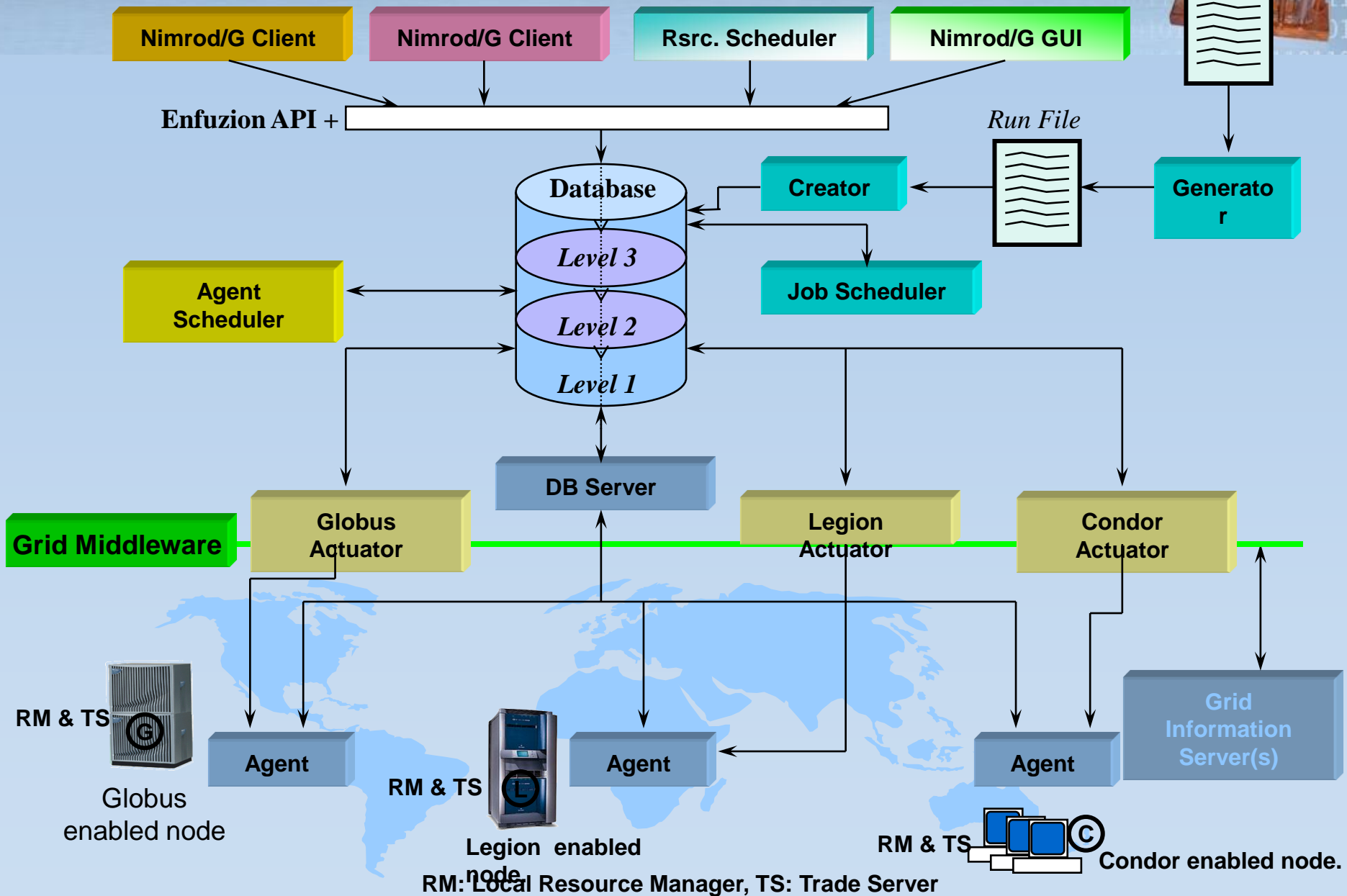
Done

Multi-Task Computing

From drug to aircraft to antenna design



Nimrod/G Architecture



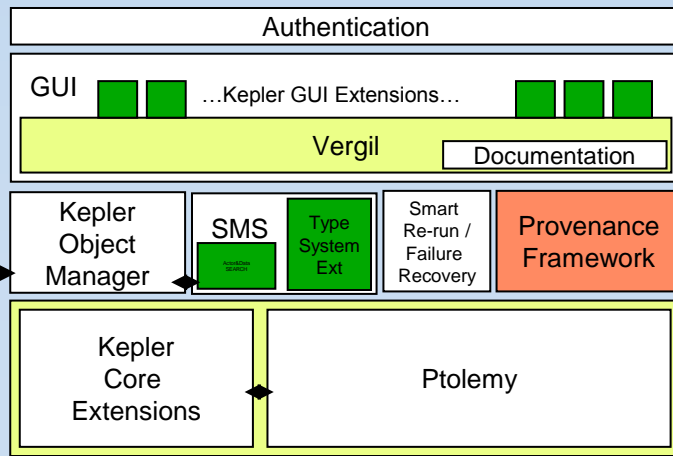
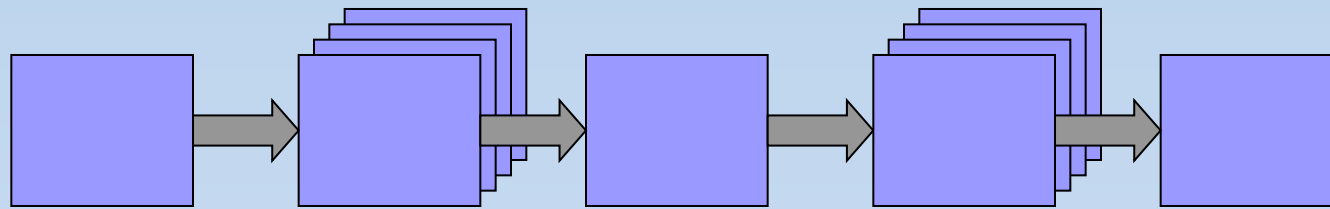


Nimrod/K Workflows

Nimrod/K Workflows



- Nimrod/K integrates Kepler with
 - Massively parallel execution mechanism
 - Special purpose function of Nimrod/G/O/E
 - General purpose workflows from Kepler
 - Flexible IO model: Streams to files



Kepler Directors



- Orchestrate Workflow
- Synchronous & Dynamic Data Flow
 - Consumer actors not started until producer completes
- Process Networks
 - All actors execute concurrently
- IO modes produce different performance results
- Existing directors don't support multiple instances of actors.

Workflow Threading



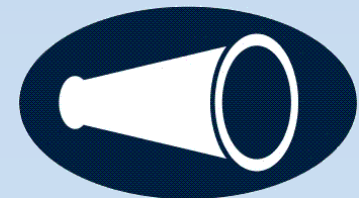
- Nimrod parameter combinations can be viewed as threads
- Multi-threaded workflows allow independent sequences in a workflow to run concurrently
 - This might be the whole workflow, or part of the workflow
- Tokens in different threads do not interact with each other in the workflow

The Nimrod/K director

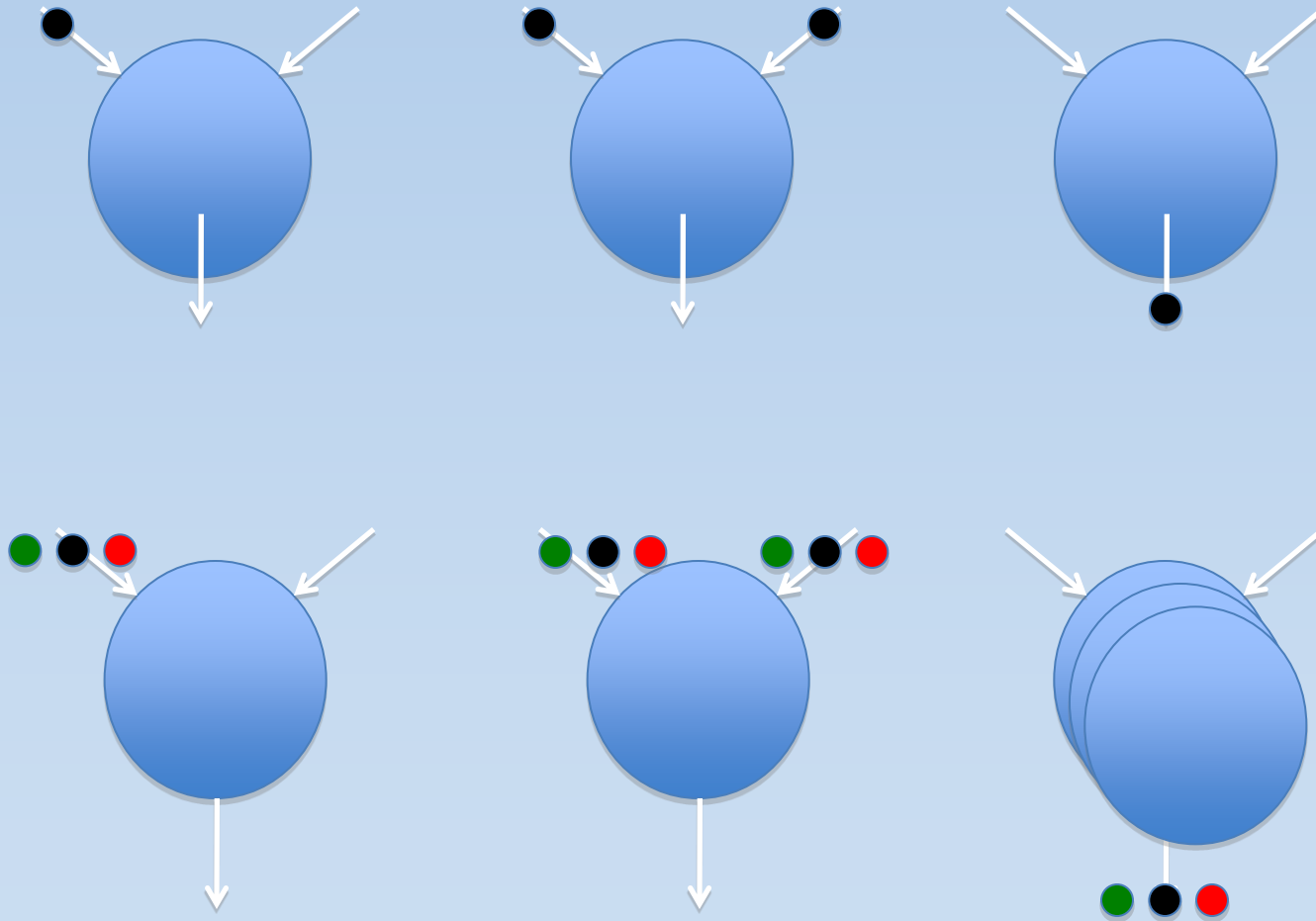


- Implements the Tagged Data Architecture
- Provides threading
- Maintains copies (clones) of actors
- Maintains token tags
- Schedules actor's events

Nimrod Director

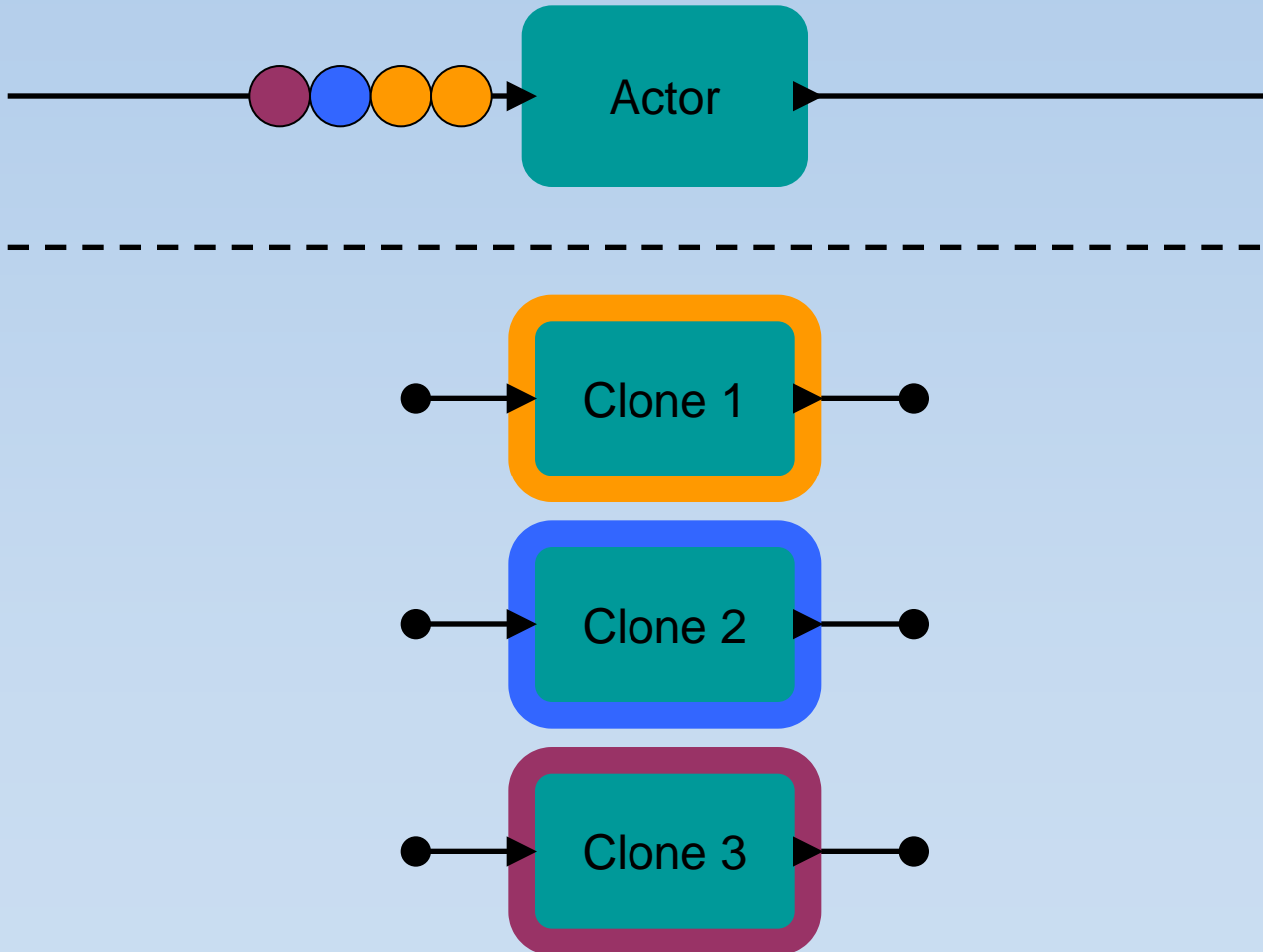


MTC through Data Flow Execution

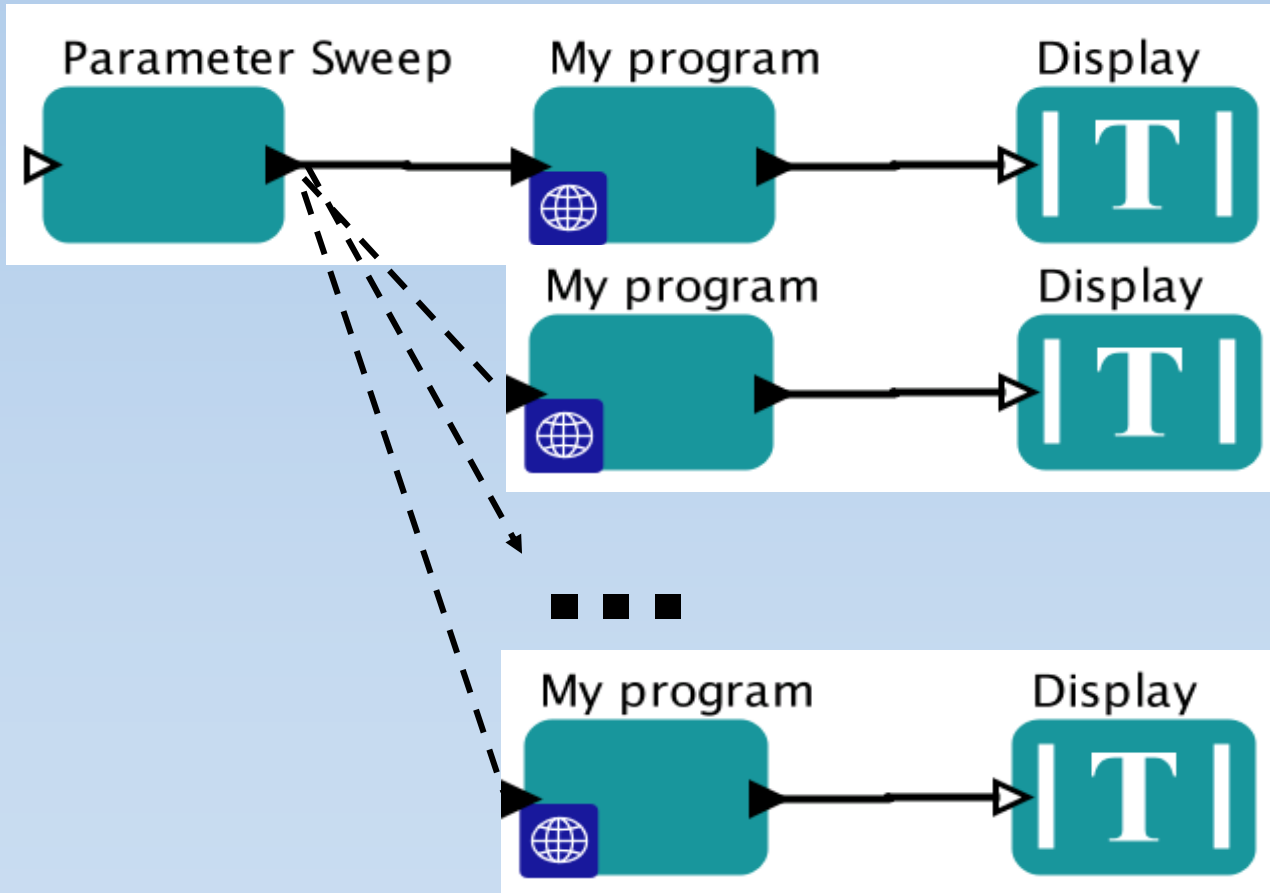


Dynamic Parallelism

Token Colouring



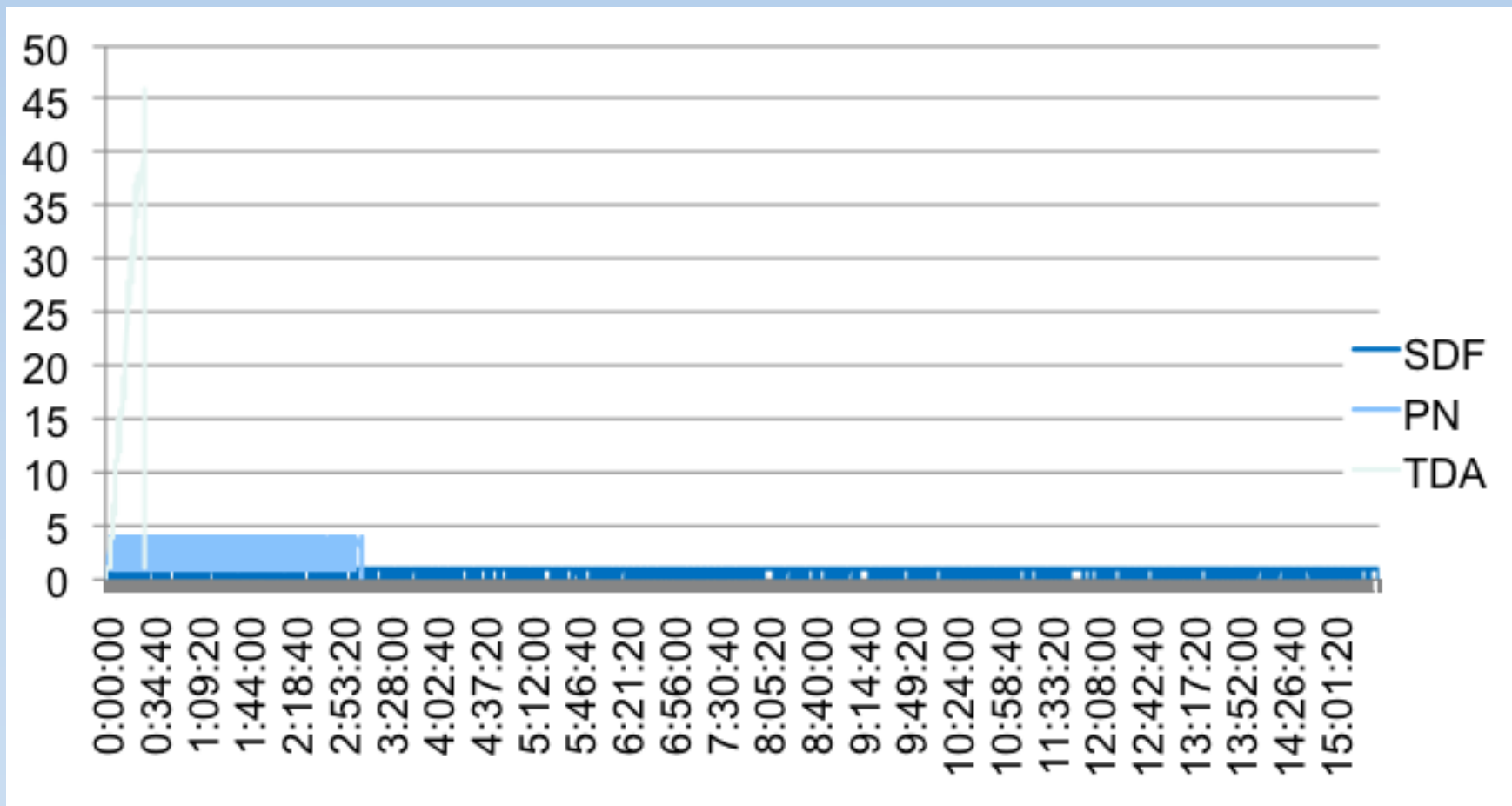
So ...



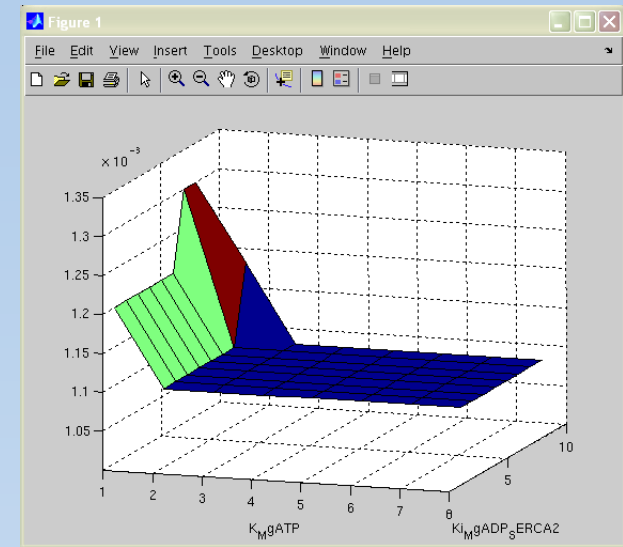
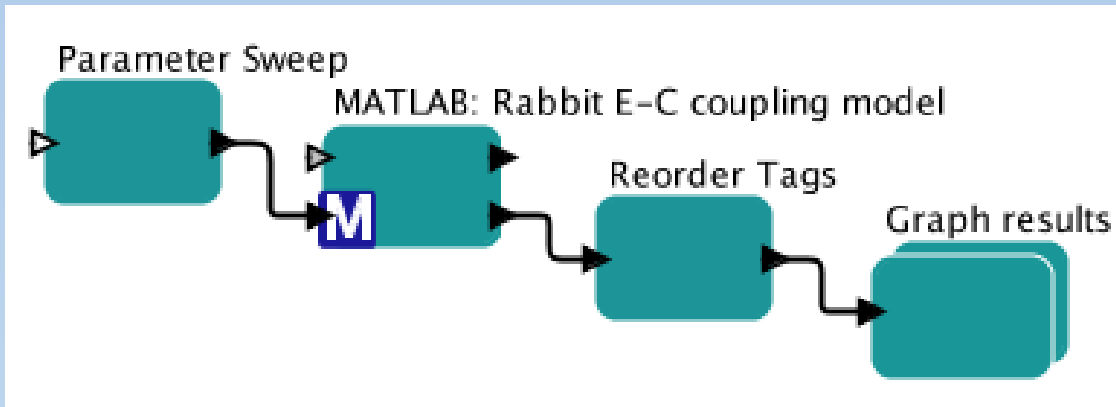
Director controls parallelism



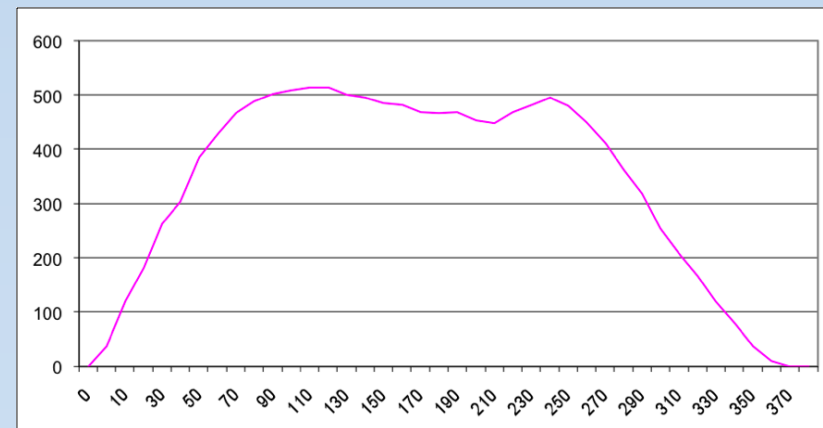
- Uses Nimrod to perform the execution



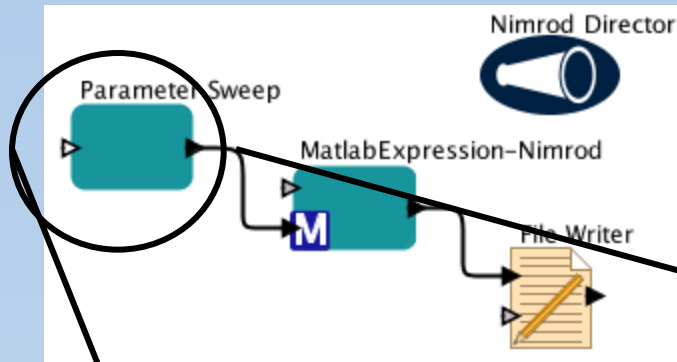
Complete Parameter Sweep



- Using a MATLAB actor provided by Kepler
- Local spawn
 - Multiple thread ran concurrently on a computer with 8 cores (2 x quads)
 - Workflow execution was just under 8 times faster
- Remote Spawn
 - 100's – 1000's of remote processes



Parameter Sweep Actor



Edit parameters for Parameter Sweep

? Sweep:

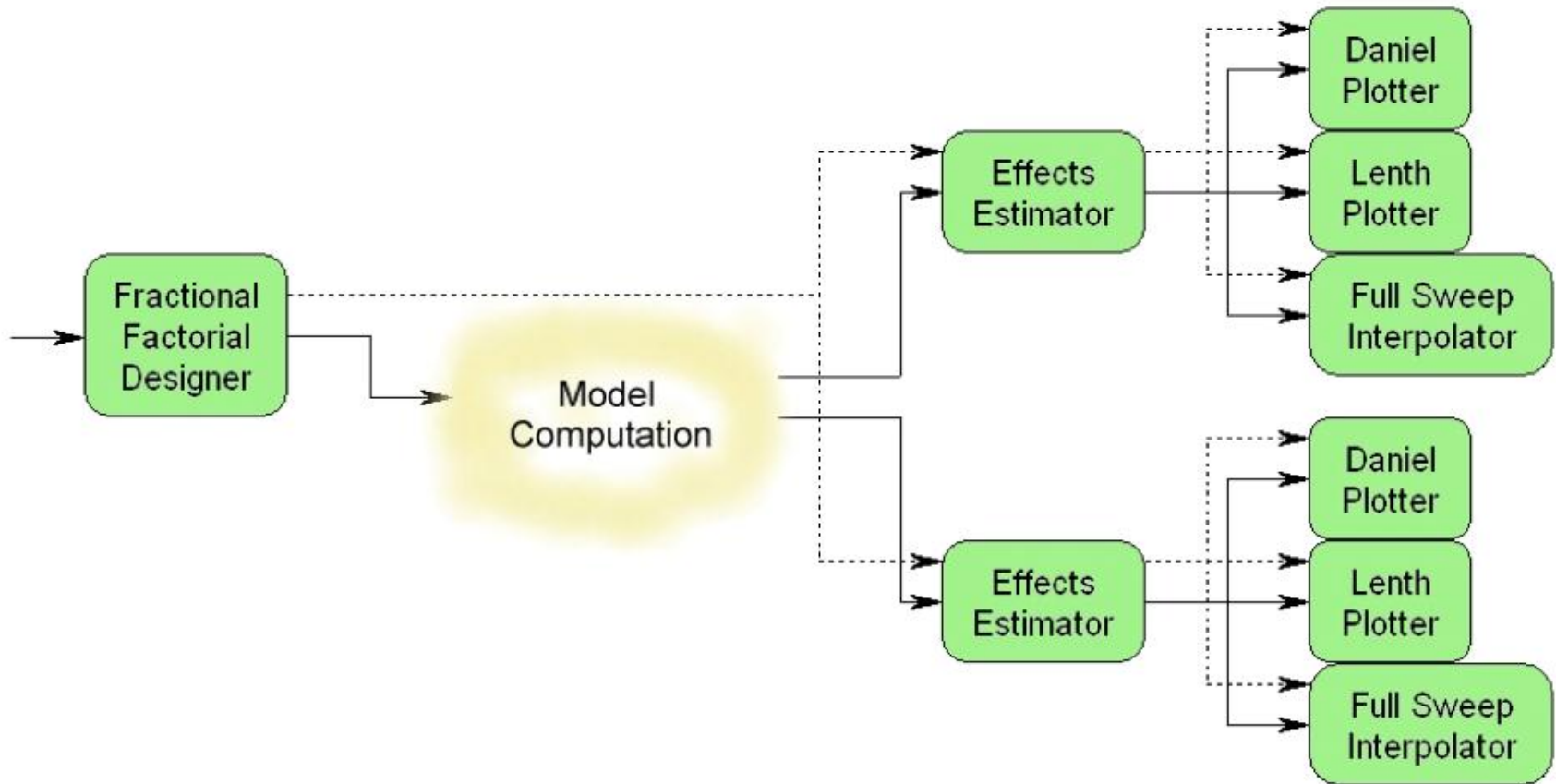
```
parameter K_MT_Nak type float range from 0.0001 to 0.008 points 2;  
parameter K_MD_Nak type float range from 0.01 to 1 points 2;  
parameter K_MgATP type float range from 1 to 2 points 2;  
parameter K_MgATP_pca1 type float range from 0.0012 to 0.12 points 2;  
parameter Ki_MgADP_pca type float range from 0.1 to 10 points 2;  
parameter K_MgATP_pca2 type float range from 0.023 to 2.3 points 2;  
parameter K_MgATP_SERCA type float range from 0.001 to 0.1 points 2;  
parameter Ki_MgADP_SERCA1 type float range from 0.014 to 1.4 points 2;  
parameter Ki_MgADP_SERCA2 type float range from 5.1 to 100 points 2;
```

class: org.monash.nimrod.ParameterSweep

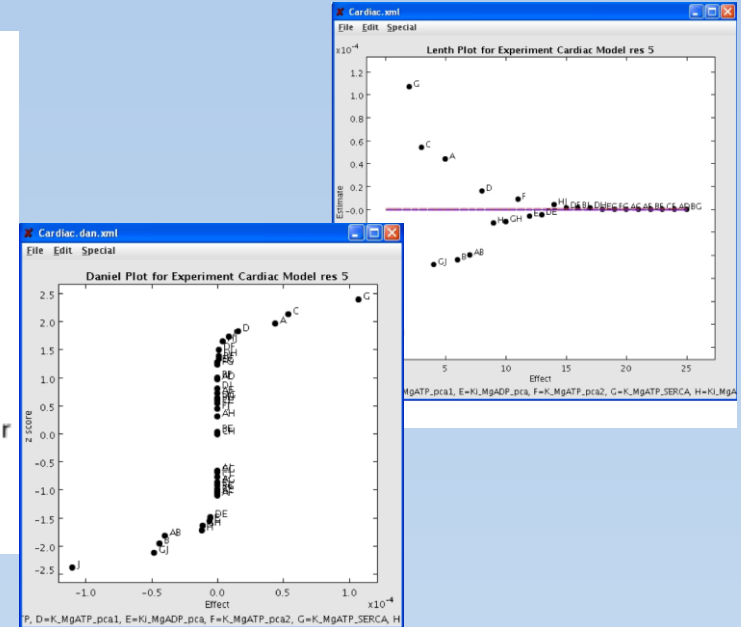
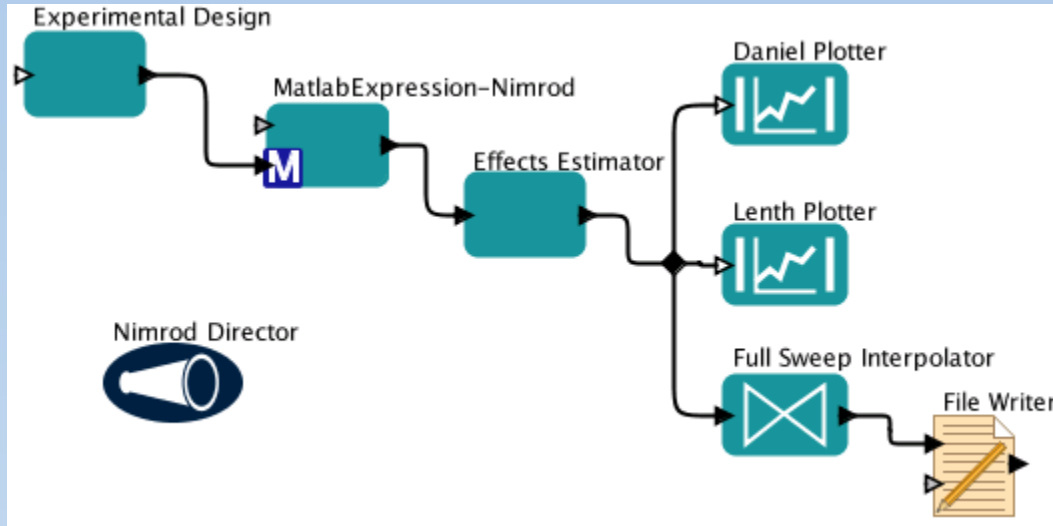
semanticType00: urn:lsid:localhost:onto:2:1#GridFunction

Commit Add Remove Restore Defaults Preferences Help Cancel

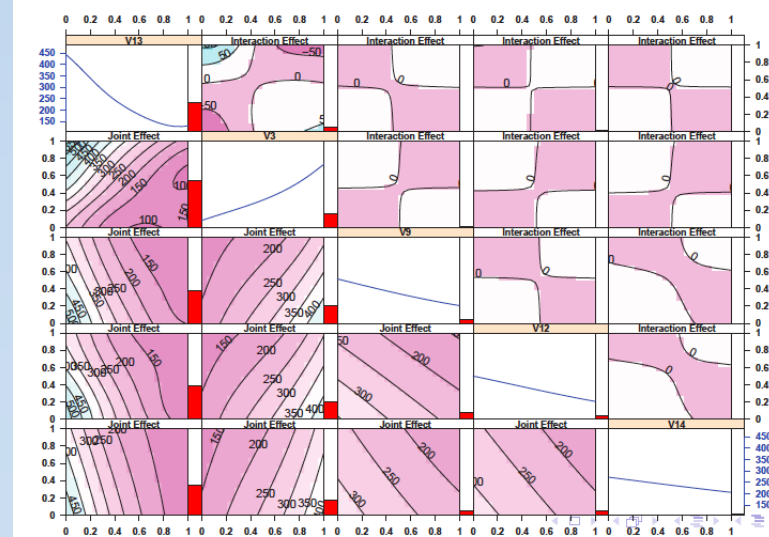
Partial Parameter Sweep



Nimrod/EK Actors

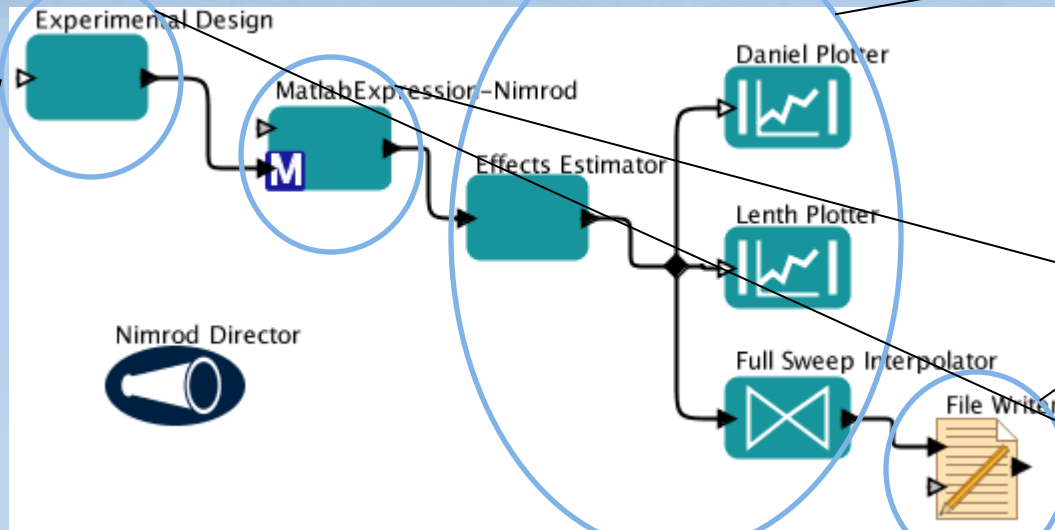


- Actors for generating and analyzing designs
- Leverage concurrent infrastructure



Nimrod/E Actors

No actor parameters need setting



No difference from the parameter sweep actors

Edit parameters for Experimental Design

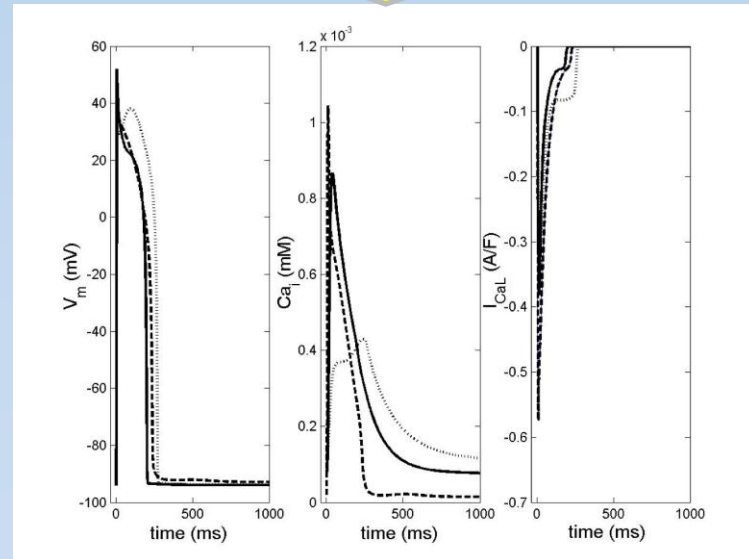
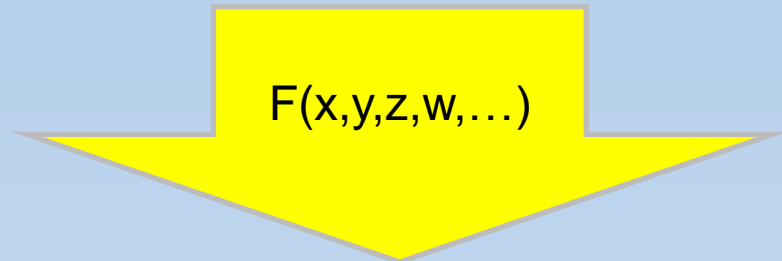
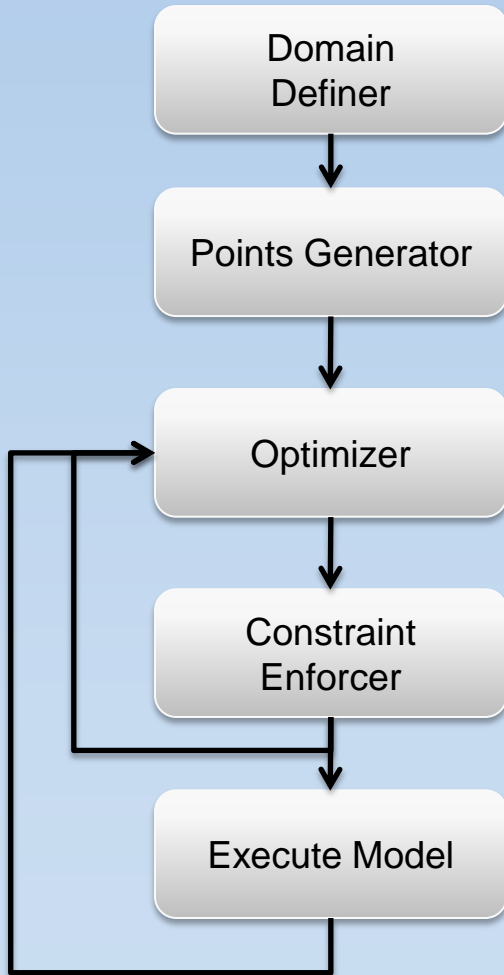
Parameters:

```
parameter Ki_MgADP_pca type float range from 0.1 to 10 points 2;  
parameter K_MgATP_pca2 type float range from 0.023 to 2.3 points 2;  
parameter K_MgATP_SERCA type float range from 0.001 to 0.1 points 2;  
parameter Ki_MgADP_SERCA1 type float range from 0.014 to 1.4 points 2;  
parameter Ki_MgADP_SERCA2 type float range from 5.1 to 100 points 2;  
  
design  
  resolution 5  
enddesign
```

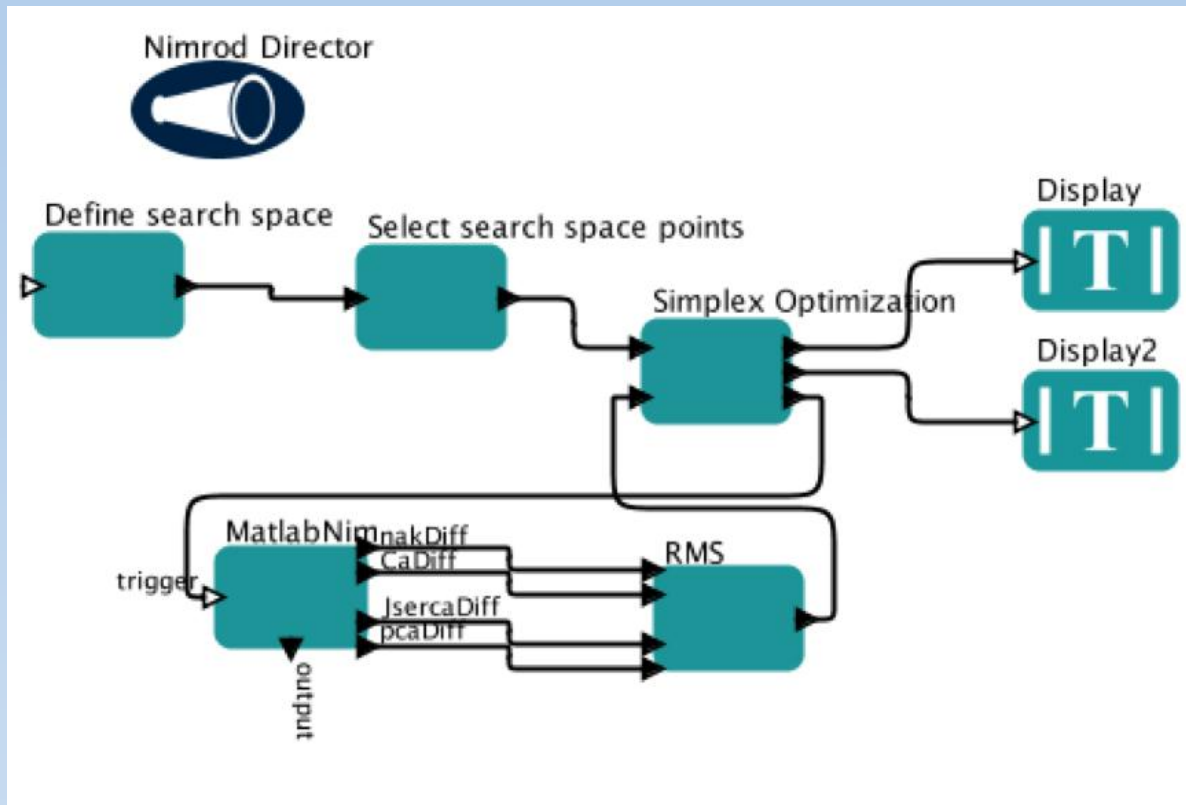
class: org.monash.nimrod.NimrodED
semanticType00: urn:lsid:localhost:onto:2:1#GridFunction

Commit Add Remove Restore Defaults Preferences Help Cancel

Parameter Optimization: Inverse Problems



Nimrod/OK Workflows



- Nimrod/K supports parallel execution
- General template for search
 - Built from key components
- Can mix and match optimization algorithms



Things the Grid ignored

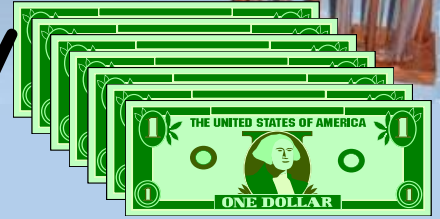
Resource Scheduling



- What's so hard about scheduling parameter studies?
 - User has deadline
 - Grid resources unpredictable
 - Machine load may change at any time
 - Multiple machine queues
 - No central scheduler
- Soft real time problem

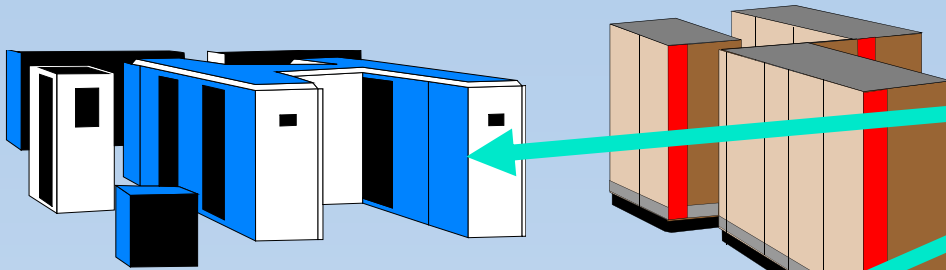


Computational Economy

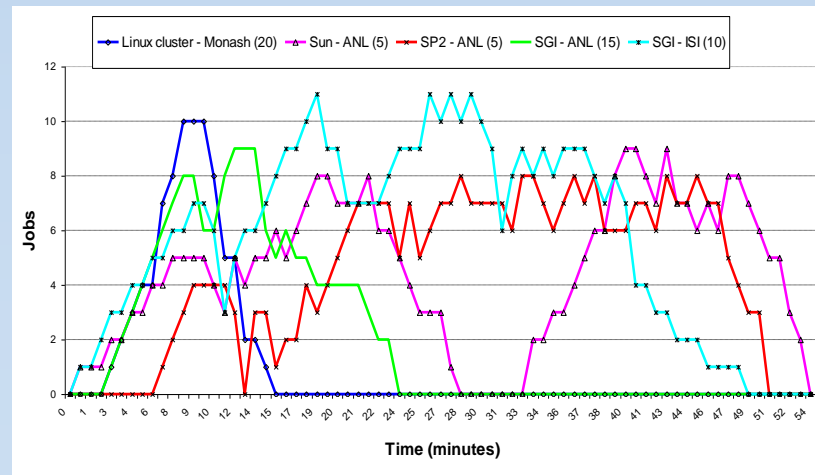
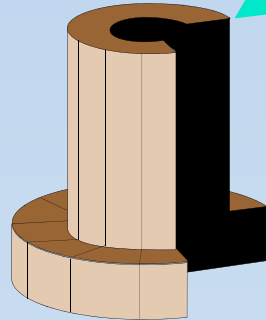
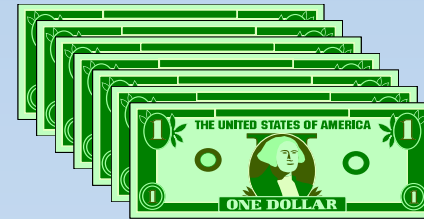


- Without cost ANY shared system becomes un-manageable
- Resource selection on based pseudo money and market based forces
- A large number of sellers and buyers (resources may be dedicated/shared)
- Negotiation: tenders/bids and select those offers meet the requirement
- Trading and Advance Resource Reservation
- Schedule computations on those resources that meet all requirements

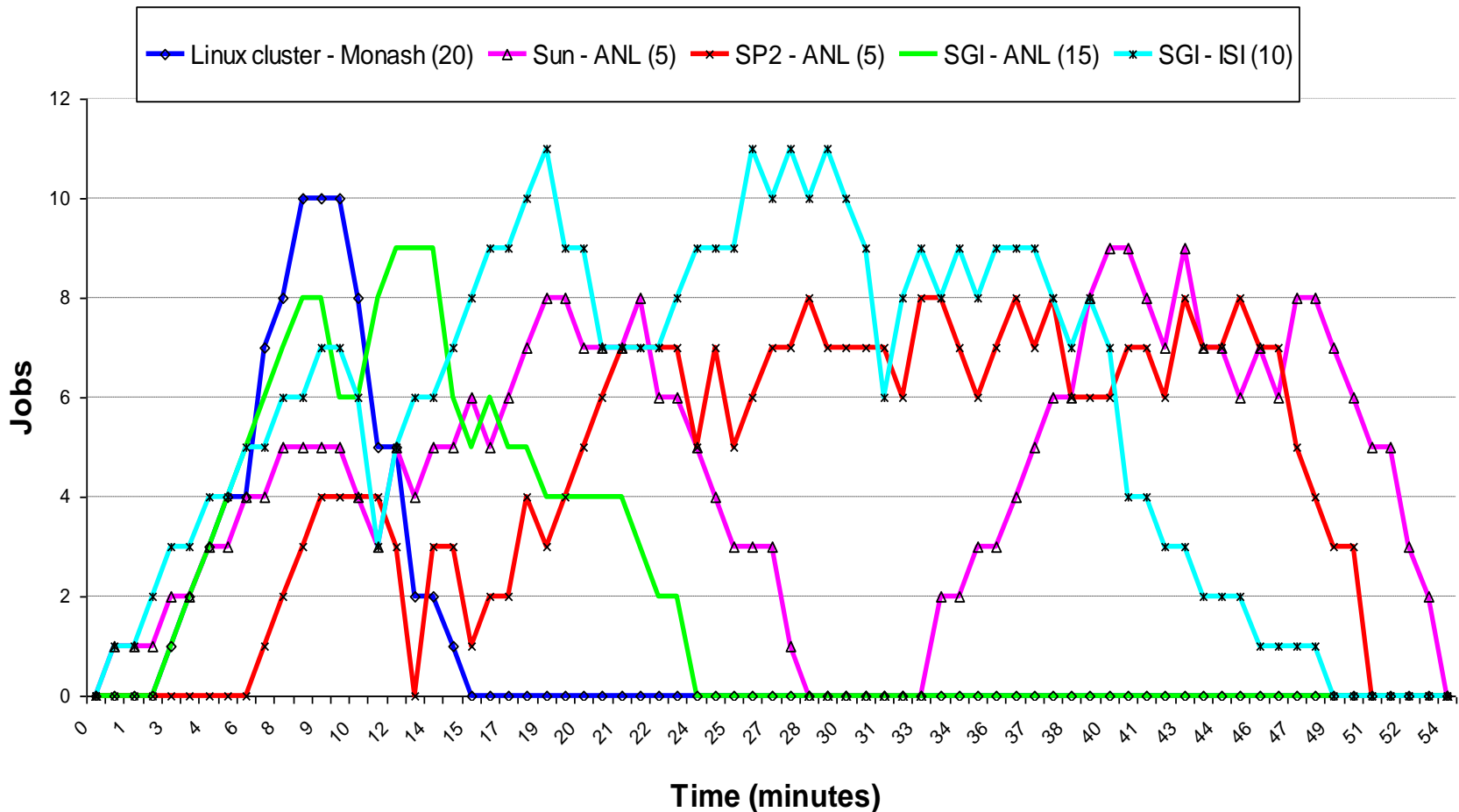
Nimrod's Scheduler



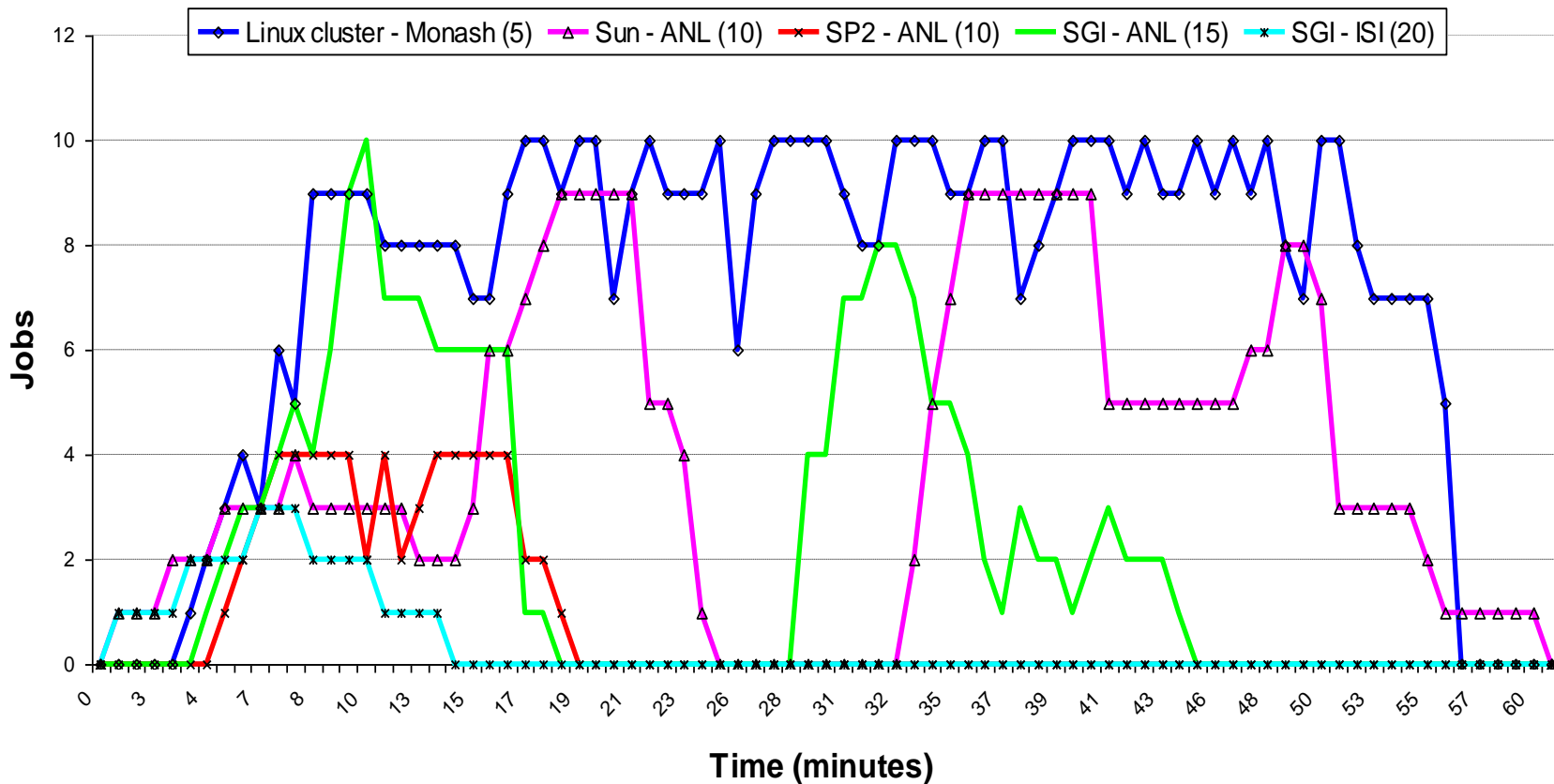
Soft real-time scheduling problem



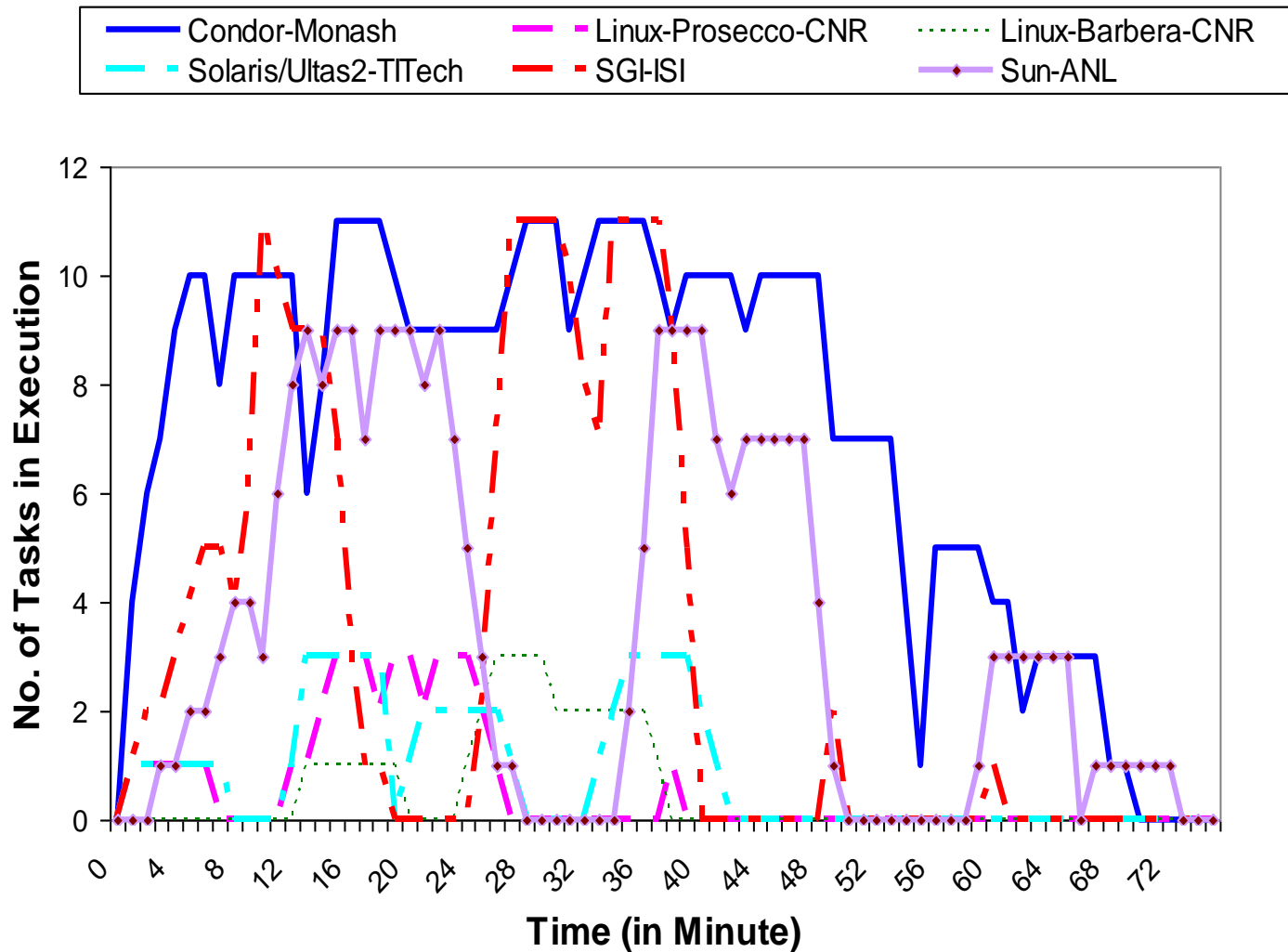
Execution @ AU Peak Time



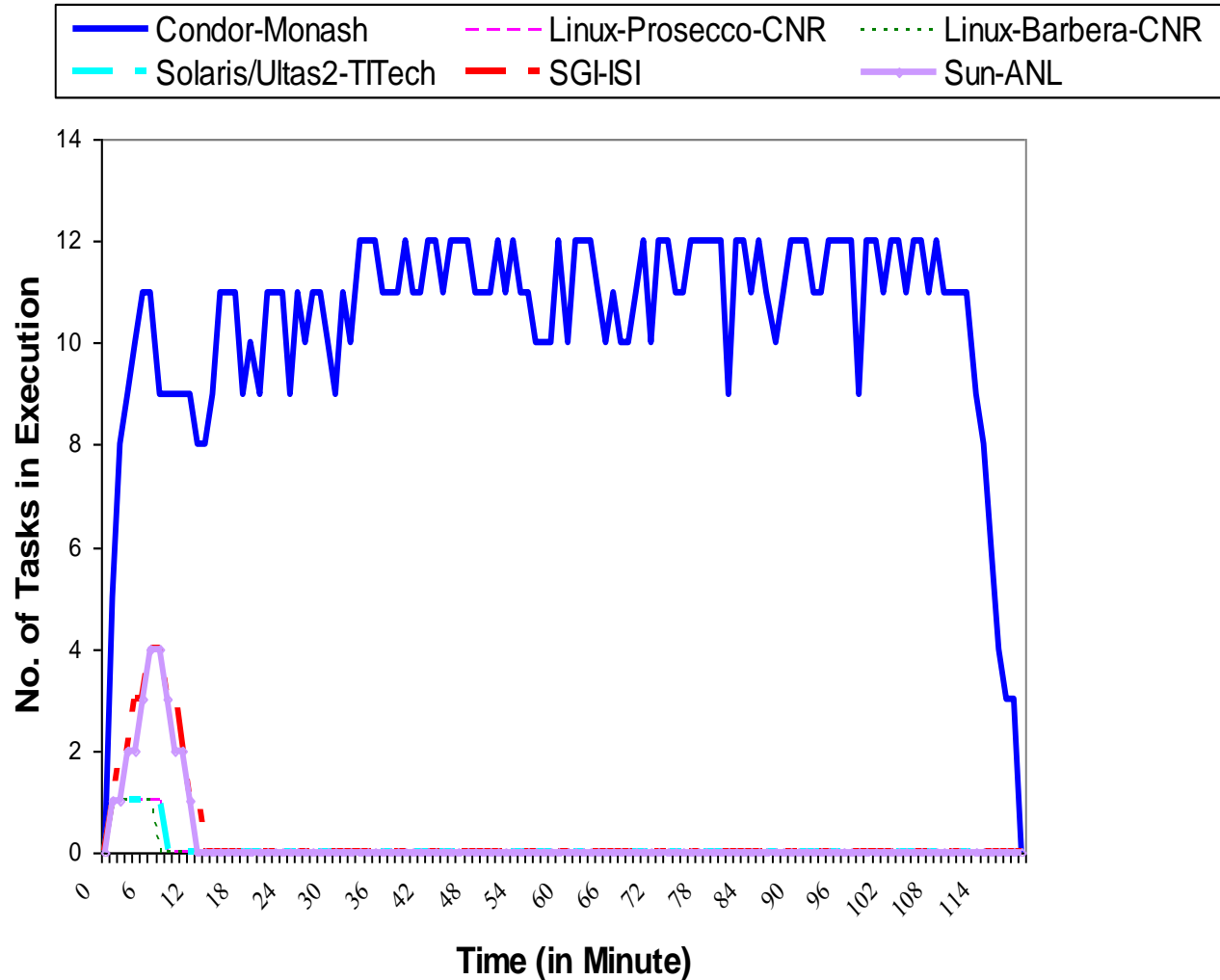
Execution @ AU Offpeak Time



Scheduling for Time Optimization



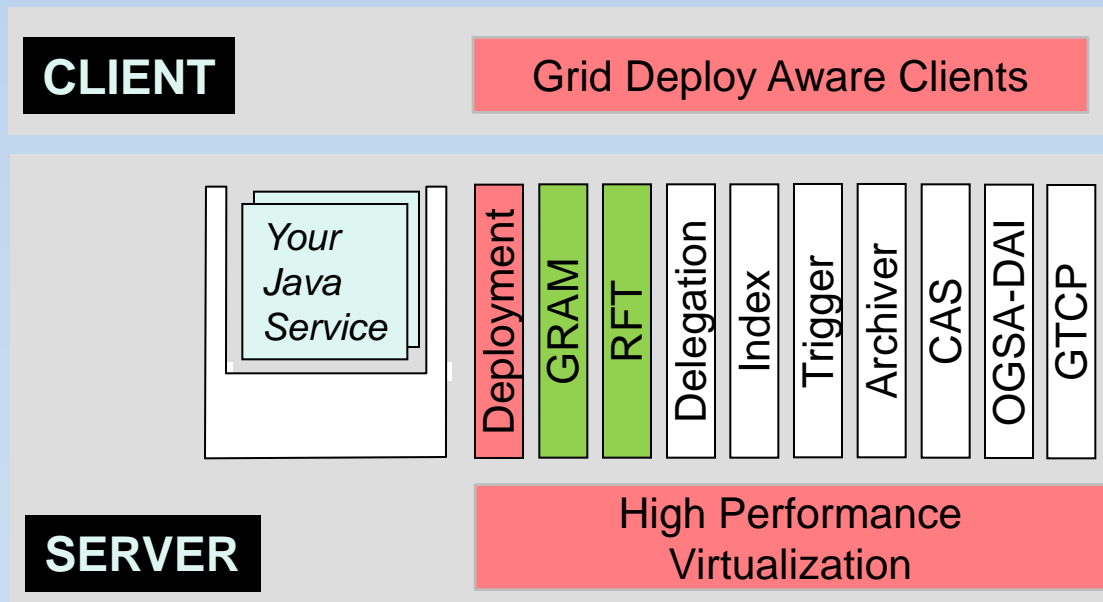
Scheduling for Cost Optimization



Deployment



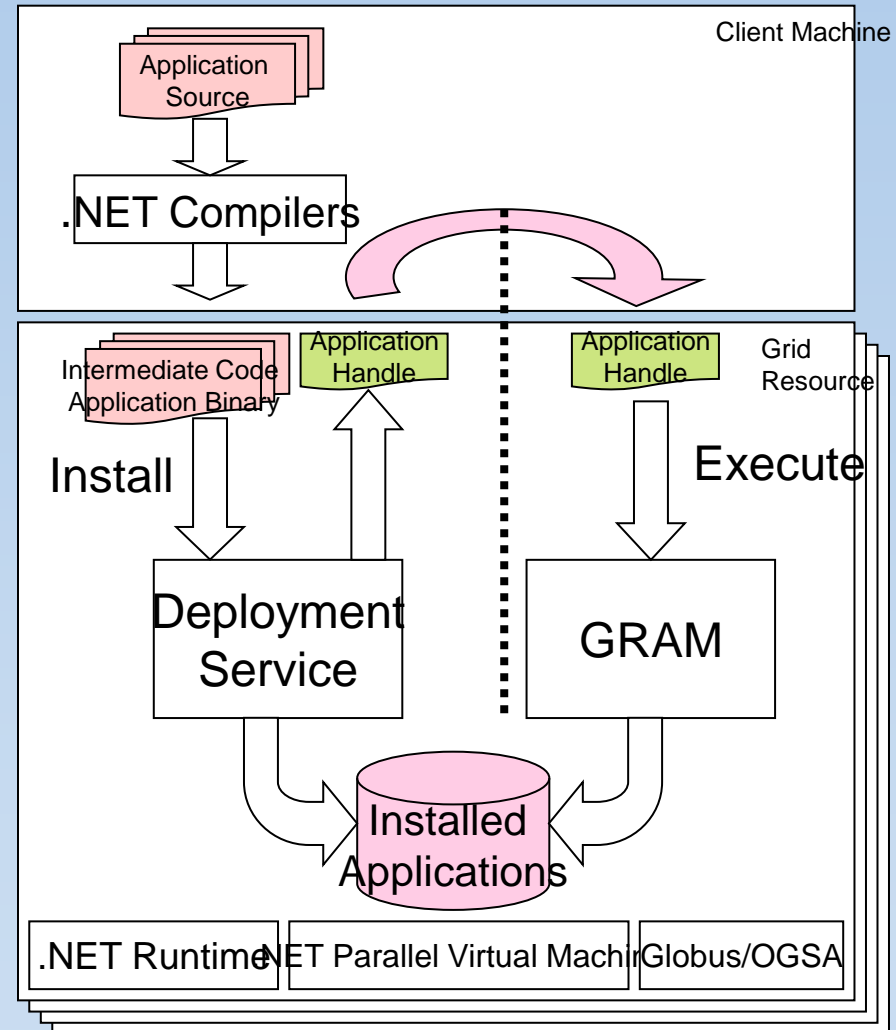
- Has largely been ignored in Grid middleware
 - Globus supports file transport, execution, data access
- Challenges
 - Deployment interfaces lacking
 - Heterogeneity



Deployment Service

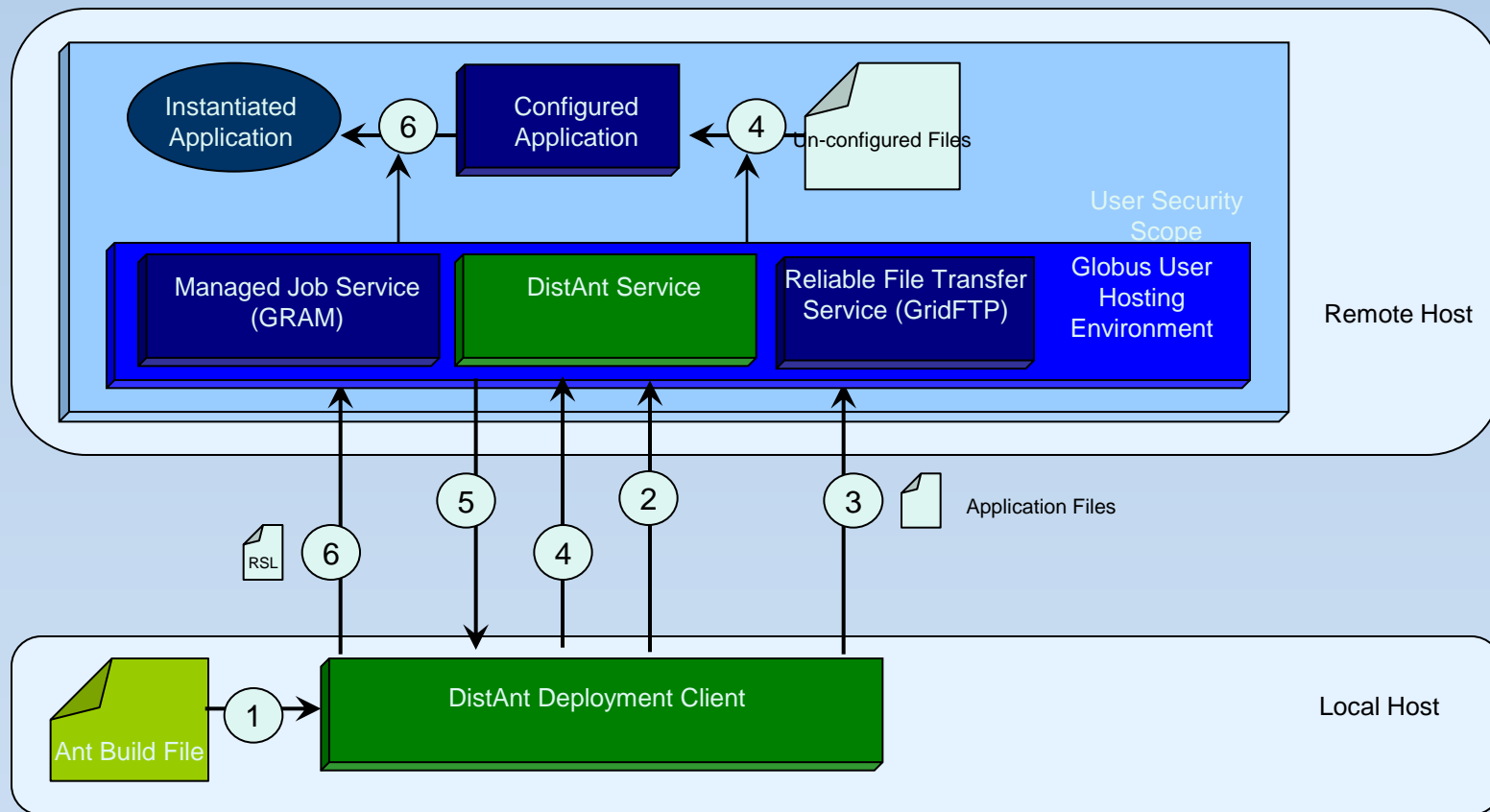


- Hide the complexity in installing software on a remote resource.
- Use local knowledge about
 - the instruction set,
 - machine structure,
 - file system,
 - I/O system, and
 - installed libraries



Towards a Grid Deployment Service

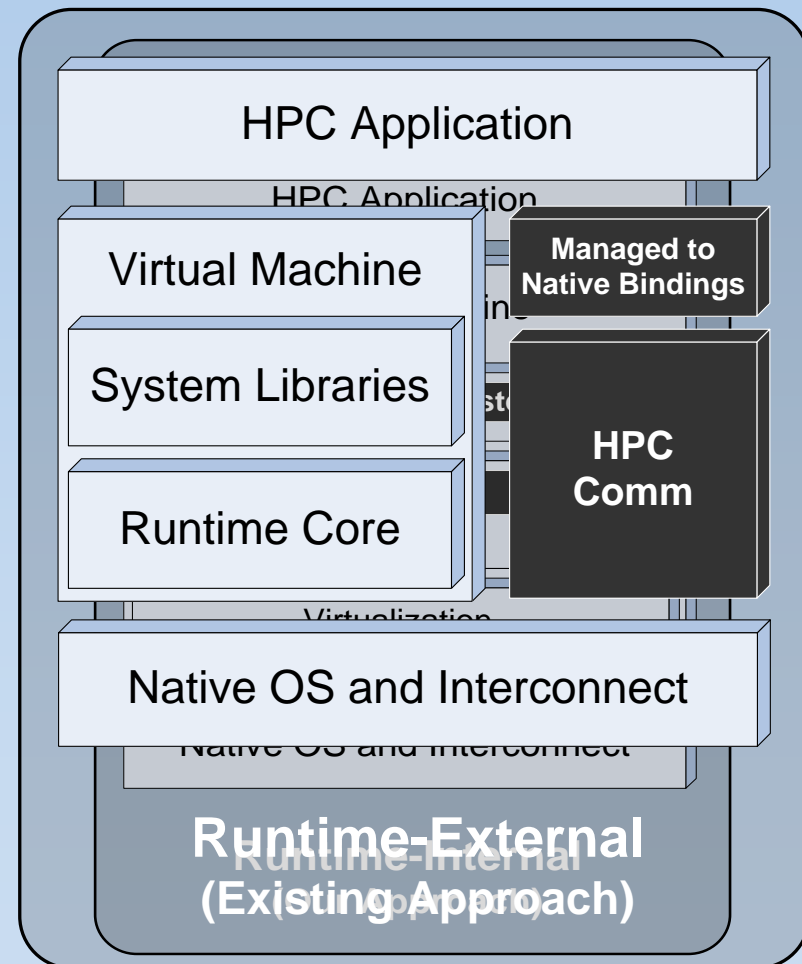
(Wojtek Goscinski)



High Performance Virtualization: The Motor Runtime



- Our approach is runtime-internal
- Why do Java & .NET support web services, UI, security and other libraries as part of the standard environment?
- Functionality is guaranteed
- Similarly, we aim to provide guaranteed HPC functionality





Clusters & Grids & Clouds

Nimrod over Clusters



Jobs / Nimrod experiment



Nimrod

Actuator, e.g., SGE, PBS, LSF, Condor



Local Batch System

axceleon™
high-performance distributed computing software

HOME | CORPORATE | PRODUCTS | SOLUTIONS | PARTNERS | DOWNLOADS | SUPPORT | NEWS | CONTACT

EnFuzion CLOUDFUZION

- Why EnFuzion
- EnFuzion Features
- EnFuzion Benefits
- Technical Specs

- User Interface
- EnFuzion Data Sheet
- EnFuzion3D Data Sheet
- CloudFuzion Data Sheet
- Applications Notes
- Demo Applications

DOWNLOAD EnFuzion

DOWNLOAD CloudFuzion

Nimrod over Grids



- Advantages

- Wide area elastic computing

- Portal based point-of-presence independent of location of computational resources

- Grid level security

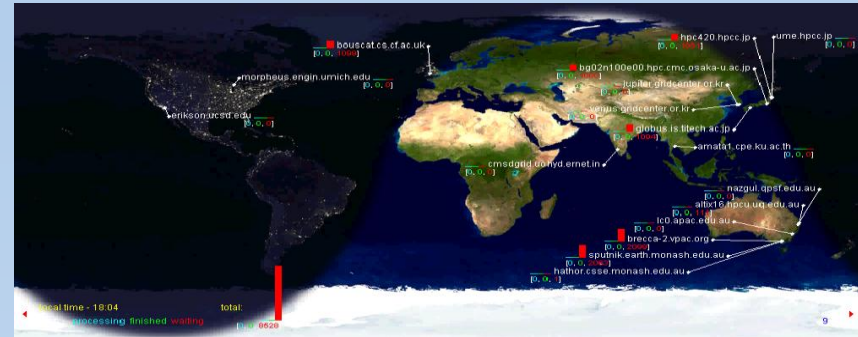
- Computational economy proposed

- New scheduling and data challenges

- Virtualization proposed (Based on .NET!)

- Leveraged Grid middleware

- Globus, Legion, ad-hoc standards





Leveraging Cloud Infrastructure

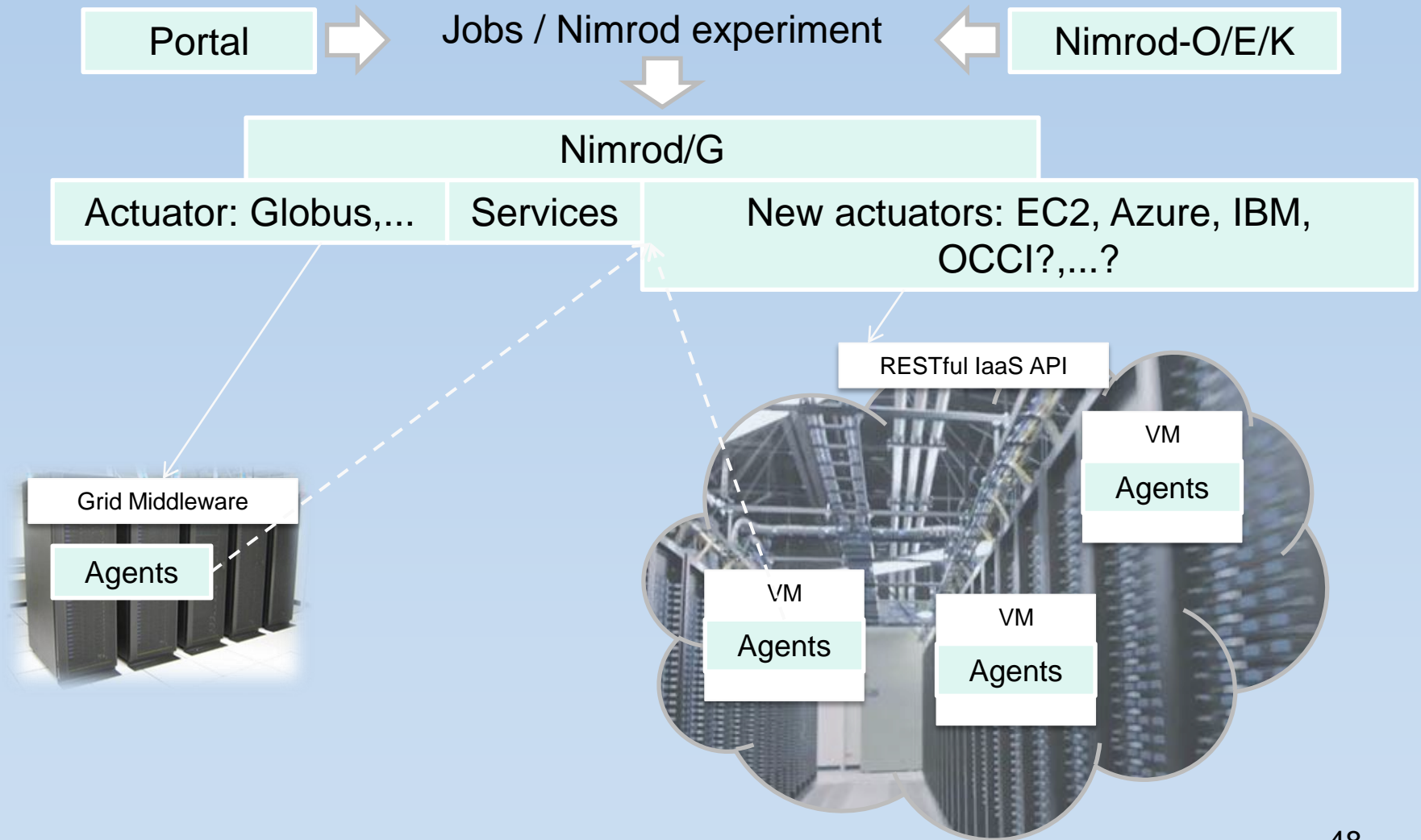
- Centralisation is easier
 - (Clusters vs Grid)
- Virtualisation improves interoperability and scalability
 - Build once, run everywhere
- Computational economy, for real
 - Deadline driven
 - "I need this finished by Monday morning!"
 - Budget driven
 - "Here's my credit card, do this as quickly and cheaply as possible."
- Cloud bursting
 - Scale-out to supplement locally and nationally available resources

Cloud Architectures



- IaaS
 - Build a virtual cluster
- PaaS
 - Leverage platform services
- SaaS
 - Nimrod portal installed on cloud

Integrating Nimrod with IaaS

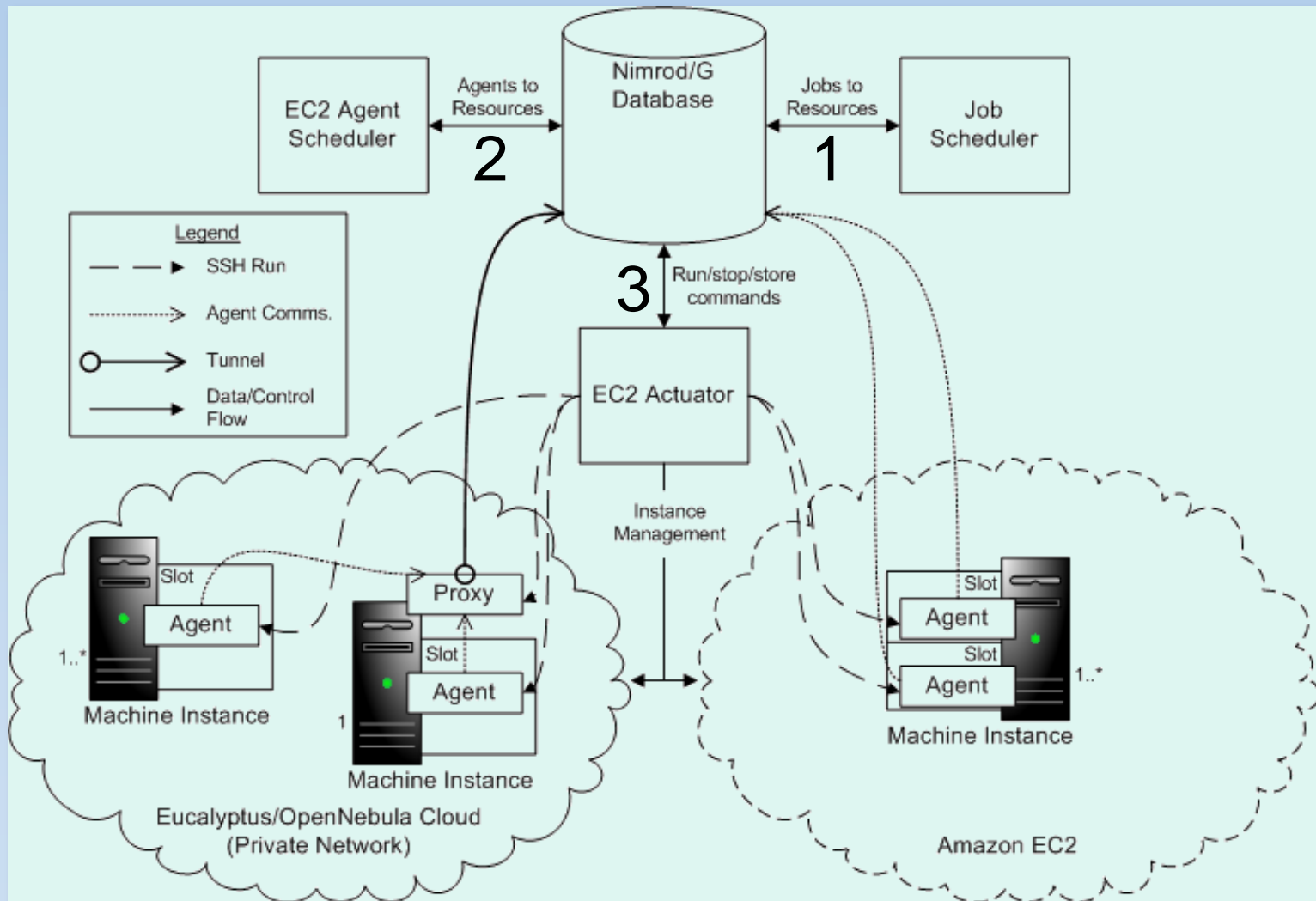


Integrating Nimrod with IaaS



- Nimrod is already a meta-scheduler
 - Creates an ad-hoc grid dynamically overlaying the available resource pool
 - Don't need all the Grid bells and whistles to stand-up a resource pool under Nimrod, just need to launch our code
- Requires explicit management of infrastructure
- Extra level of scheduling - when to initialise infrastructure?

Integrating Nimrod with IaaS



PaaS is trickier...



- More variety
 - Azure vs AppEngine
- Designed for web-app hosting
 - Nimrod provides a generic execution framework
- Higher level PaaS too prescriptive
 - AppEngine: Python and Java only

Nimrod-Azure Mk.1



- Nimrod server runs on a Linux box external to Azure
- Nimrod-Azure actuator module contains the code for managing Nimrod agents on Azure
 - pre-defined minimal NimrodWorkerService cspkg;
 - library for speaking XML over HTTP with the Azure Storage and Management REST APIs

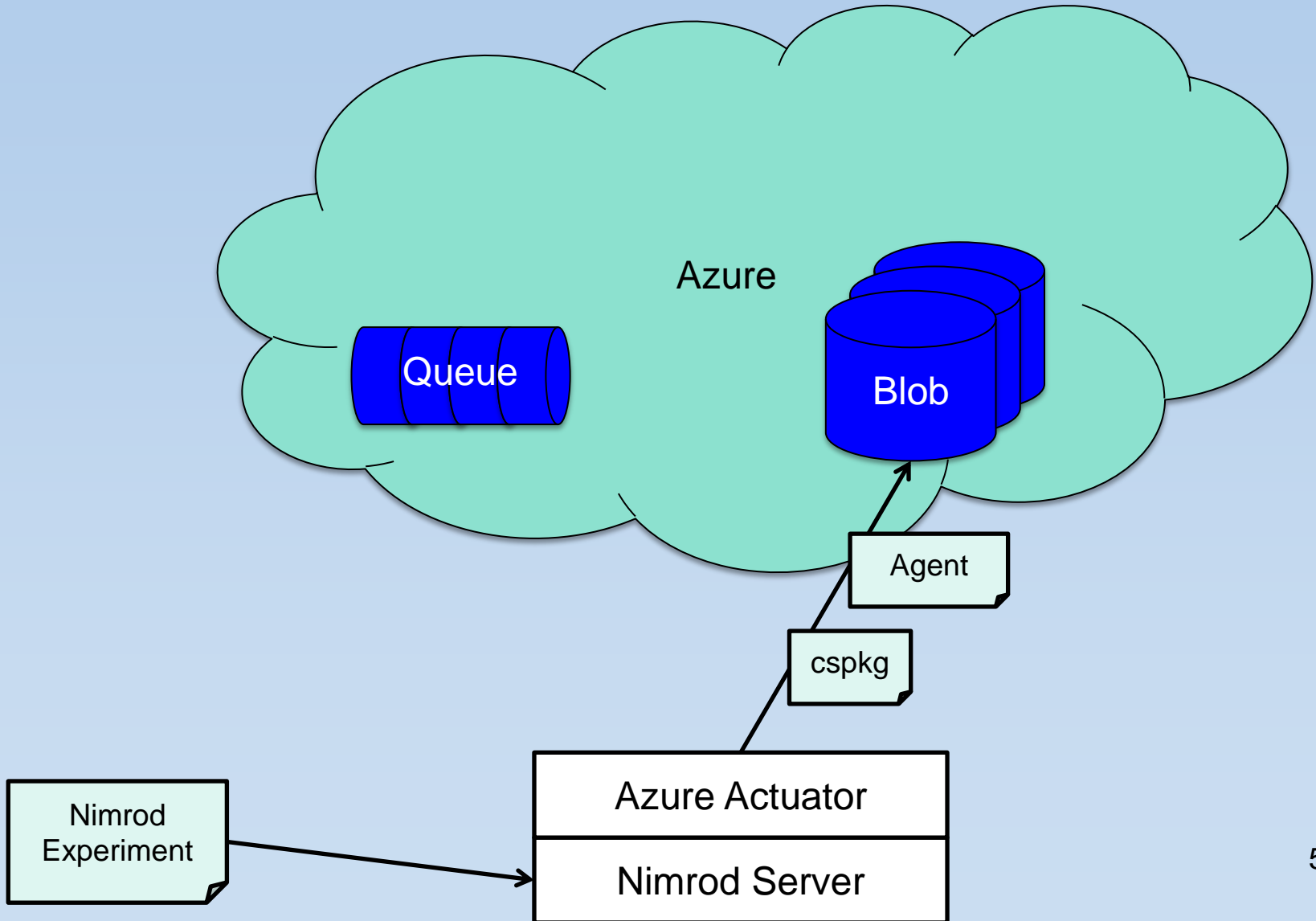
Integrating Nimrod with Azure



To stand-up an Azure compute resource under Nimrod, the *actuator*.

- Copies the Nimrod *agent* package and encryption keys to an Azure Blob
- Adds command line parameters for *agents* to an Azure Queue
- Builds an initial cscfg for the deployment including relevant blob and queue URLs
- Deploys the service to the Cloud

Integrating Nimrod with Azure



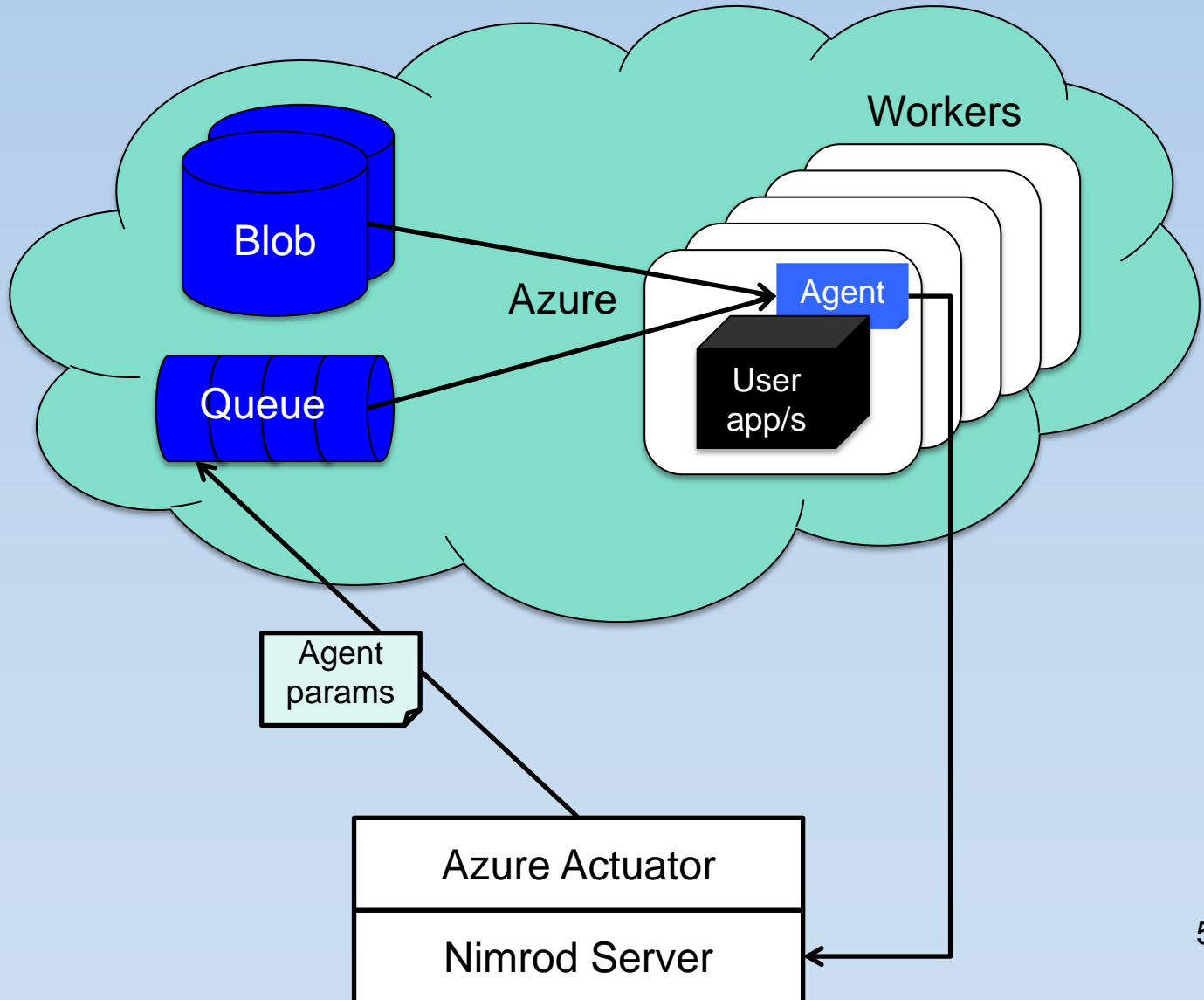
Integrating Nimrod with Azure



Once deployed, the NimrodWorkerService:

- Pulls the Nimrod *agent* package from blobs referenced in cscfg settings
- Unpacks and launches the *agent* with parameters from the queue referenced by cscfg
- The *agent* connects out to the Nimrod server, pulling work and pushing results until: no work left; lifetime ends; exception
- But, when the *agent* exits there is no way to de-provision the role instance... scaling without de-scaling?! Please fix this!

Integrating Nimrod with Azure



Grid + Amazon + Azure



Status of EC2vAzure1

The experiment had executed for 1hr, 54mins and 44secs.

All 8192 jobs have completed.

[Check for grid errors](#)

[Archive or Delete the Experiment](#) [Reset Experiment](#)



Plan file

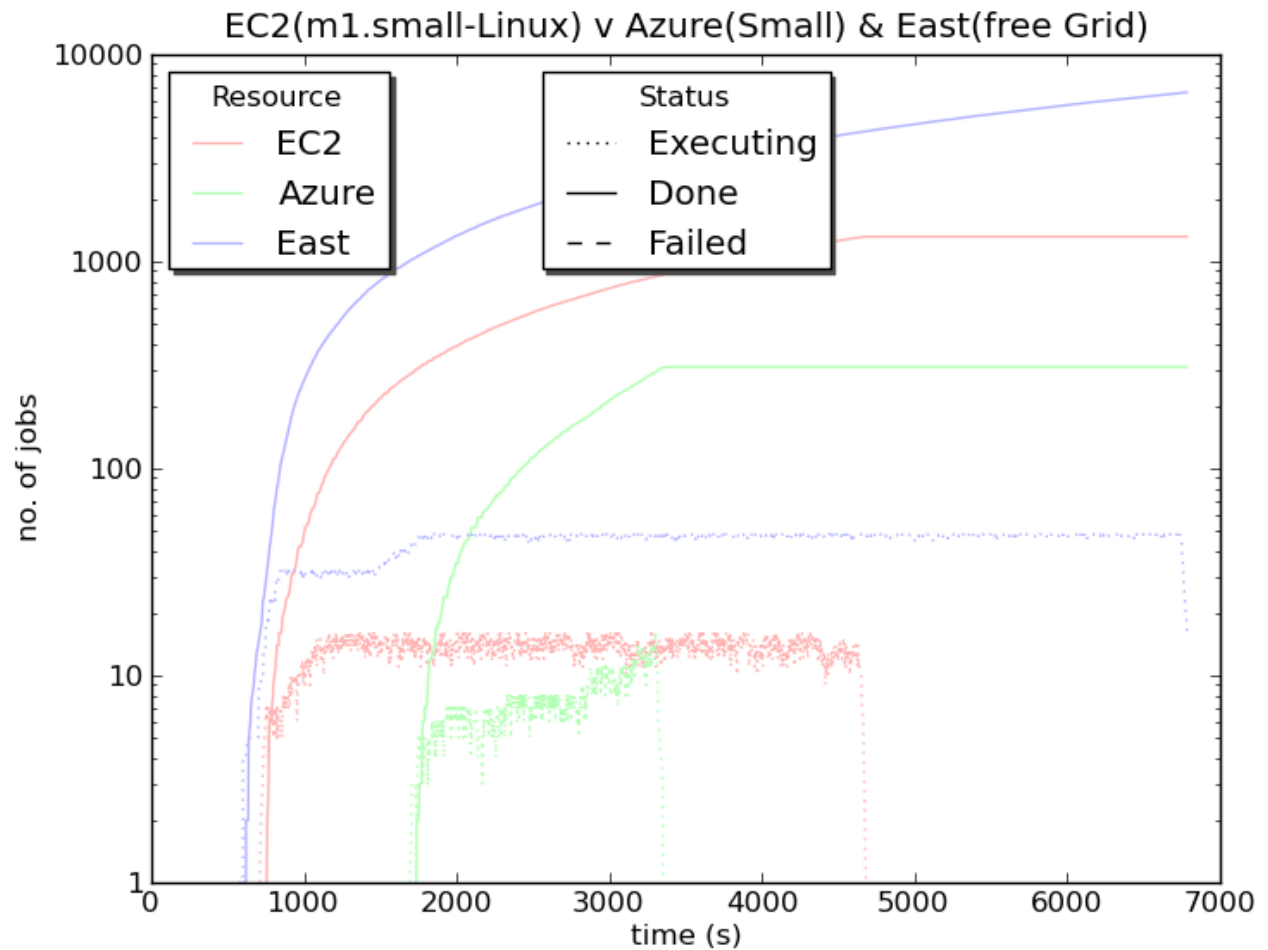
```
parameter time integer random from 20 to 40;
parameter jobs integer range from 1 to 8192 step 1;
```

task nodestart

```
copy input node:.;
copy magic.bat.$OS node:magic.bat.skel
endtask
```

task main

```
node:substitute magic.bat.skel magic.bat
onerror ignore
node:execute chmod +x magic.bat
onerror fail
node:execute magic.bat > out
node:execute hostname >> out
copy node:out out/out.$jobname
endtask
```



Conclusions and Future Directions



- Commercial Clouds
 - Grid economy == commercial clouds
 - Virtualisation built into fabric
- Leverage MTC paradigm
 - More complex Design of Experiments
 - More optimization Algorithms
- Make environment more useful
 - New portal
 - Workflows that interact with IO devices and Portals

Questions?



More information:

<http://messagelab.monash.edu.au>

Acknowledgements

MeSsAGE Lab



- Faculty Members
 - Jeff Tan
 - Maria Indrawan
- Research Fellows
 - Blair Bethwaite
 - Slavisa Garic
 - Jin Chao
- Admin
 - Rob Gray
- Current PhD Students
 - Shahaan Ayyub
 - Philip Chan
 - Colin Enticott
 - ABM Russell
 - Steve Quinette
 - Ngoc Dinh (Minh)
- Completed PhD Students
 - Greg Watson
 - Rajkumar Buyya
 - Andrew Lewis
 - Nam Tran
 - Wojtek Goscinski
 - Aaron Searle
 - Tim Ho
 - Donny Kurniawan
 - Tirath Ramdas
- Funding & Support
 - **Amazon**
 - Axceleon
 - Australian Partnership for Advanced Computing (APAC)
 - **Australian Research Council**
 - Cray Inc
 - CRC for Enterprise Distributed Systems (DSTC)
 - GrangeNet (DCITA)
 - Hewlett Packard
 - IBM
 - **Microsoft**
 - Sun Microsystems
 - US Department of Energy

