Workflow-based comparison of two Distributed Computing Infrastructures

J. Montagnat, T. Glatard, D. Reimert, K. Maheshwari, E. Caron, F. Deprez
(CNRS, I3S / CREATIS, INRIA, LIP)

WORKS'10
New Orleans
November 14, 2010

http://gwendia.polytech.unice.fr
• **Objectives**
  – Evaluate performance of different Distributed Computing Infrastructures (DCIs): a production (European EGI – former EGEE) and a research (French G5K) infrastructure

• **Motivations**
  – Workflow-based applications can be easily ported to different DCIs (or simultaneously use different DCIs)
  – DCIs hardware and middleware significantly differ
  – Distributed computing performance is difficult to assess

• **Method**
  – Experiments-based: same workflow application executed on different DCIs
  – Execution conditions aligned as much as possible
  – Comparison criterions identification and measurement
Different DCI models

• **Infrastructures**
  - EGI: production, 250+ computing centers, 160k+ CPU cores, 10k+ users, world-scale, gLite middleware (batch-oriented)
  - G5K: research, 9 sites, 5k+ CPU cores, 100's users, national-scale, reconfigurable (any middleware), reservable resources

• **Resources usage**
  - EGI: production = permanent (yet variable) workload
    ▪ SRM-compatible storage resources
    ▪ Amount of resources never precisely known
    ▪ WAN communications
    ▪ High-end resources in well equipped computing centers
  - G5K: research = higher workload variations
    ▪ NFS access to disks
    ▪ Controlled amount of resources
    ▪ National WAN communication on a private high-performance network
    ▪ 1-5 years old resources
Different DCI models

- **Middleware**
  - EGI: gLite
    - Batch-oriented computations
    - File servers with heavy compatibility front-ends
    - Scientific Linux (REHL-like) v4 or 5 OS
  - G5K: OAR resources reservation
    - Dedicated resources, any middleware
    - NFS servers site-wise, manual data transfer across sites (scp...)
    - Any OS system image

- **Heterogeneity**
  - All IA32/64-compatible CPUs
  - Although significant hardware variations cause practical problems for OS images deployment
• **Cardiac image segmentation workflow**
  - 2 initialization stages (mhd2qc + ImgAndModelInit)
  - Multiple instances of the segmentation process (det3D4)
Workflow runs

- **Parameter sweep application (parameters-combinatorial)**
  - Small-size: 2+12 segmentation instances (testing)
  - Medium-size: 2+200 segmentation instances (scale-up)
  - Large-size: 2+2080 segmentation instances (challenging)

- **Same binaries ran on each infrastructure**
  - Binaries compiled for SL4
  - SL4 OS image installed on G5K nodes (proved to be painful!)

- **Fixed-size infrastructure**
  - 54 (= 3 x 18) cores reserved for most experiments

- **Experiments were reproduced 3 to 5 times**
  - Compensate for inter-experiments variability
  - Results are given as average value +/- standard deviation

- **Experiments were ran on a single site or on 3 sites**
  - Both intra-site and WAN communications
Quantitative comparison

Grid Workflow Efficient Enactment for Data Intensive Applications

- **Compare EGI and G5K performance in similar conditions**
  - Allocate same size infrastructure and run same workflows
  - Measure makespan, data transfer time, activities execution time and idle time
- **DIANE pilots on EGI**
  - Resource reservation
  - Pilots submitted to batch using GASW
  - Pilots may fail (faulted, expired, killed by sysadmin, unreachable...)
- **Pilots used to reserve resources**
  - Need a fixed-number pool of pilots
  - Over-provisioning to replace failed pilot without delay
  - Submission of idle pilots until the needed number is available
- **54 resources reserved for experiment runs**
  - 70 to 90 pilots submitted for each experiment
Small-size runs

- **EGI (1 site)**

- **G5K (1 and 3 sites)**

  - **Features**
    - gLite overhead
    - Resources heterogeneity on G5K (1 to 2 CPU time)
Medium-size runs

- EGI (1 site)
- EGI (3 sites)
- G5K (1 site)
- G5K (3 sites)
Medium-size runs

- **Features**
  - Batches of 54 concurrent tasks
  - Desynchronization over time
  - Input files caching
  - DIET workflow decomposition strategy

- **Few task failures on EGI**
  - Causing resubmission

- **Difference between 1 and 3 sites runs**
  - Little impact on EGI; more impact on G5K (e.g. data transfers)

- **Makespan variability is higher on G5K than on EGI**
  - No better reproducibility on G5K than on EGI using pilots
Large-size runs

- EGI (3 sites)

- EGI (failed)

- G5K (1 site)

- G5K (3 sites)
Large-size runs

- **Features**
  - Linear profile

- **Many failed experiments**
  - EGI: pilot lifetime limitations
  - G5K: difficulty to proceed with reservations and platform failure

- **Reproducibility**
  - Higher on single site than on 3 sites with EGI
  - Higher on 3 sites than on single site with G5K
• **Greedy pilots allocation**
  – No limitation to 54 pilots
  – ~30 sites
  – ~3% failures

• **Features**
  – Delayed start (time for first pilots to register)
  – Sub-linear profile (more resources available)
  – Diane's favorite heavy tail

• **Performance**
  – Comparable makespan as with controlled conditions (54 pilots)
Conclusions

• Difficulty to compare different DCIs performance

• Experiments-based performance measurement
  – Sensitive to the workflow properties (e.g. the workflow used features maximal data parallelism and no critical bottleneck activity)

• Experimental setup
  – Aligning execution conditions with pilot jobs + pilot population controller + single runtime
  – Limited in scale

• Infrastructure properties outlined
  – Difference in CPU performance, network topology and middleware
Conclusions

• **A 54-nodes controlled infrastructure reaches makespans close to EGI knowing that:**
  – Experiments on EGI have been run on large, reliable sites
    - < 5% error rate in all cases
  – EGI can handle several concurrent users and experiments
  – Few failures are highly impacting makespan in production

• **Reproducibility may be as good on EGI as on G5K under controlled condition**
  – Feasibility of large-scale experiments on EGI