SILECS/Grid’5000: le volet data-center de SILECS: Présentation et exemples d’expériences

Lucas Nussbaum

► To cite this version:

Lucas Nussbaum. SILECS/Grid’5000: le volet data-center de SILECS: Présentation et exemples d’expériences. TILECS - Towards an Infrastructure for Large-Scale Experimental Computer Science, Jul 2019, Grenoble, France. hal-02401836

HAL Id: hal-02401836
https://inria.hal.science/hal-02401836
Submitted on 10 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
SILECS/Grid’5000

Le volet data-center de SILECS
Présentation et exemples d’expériences

Lucas Nussbaum
Grid'5000 Technical Director

Workshop TILECS
2019-07-03
The Grid’5000 testbed

- A large-scale testbed for distributed computing
  - 8 sites, 31 clusters, 828 nodes, 12328 cores
  - Dedicated 10-Gbps backbone network
  - 550 users and 120 publications per year
The Grid’5000 testbed

- A large-scale testbed for distributed computing
  - 8 sites, 31 clusters, 828 nodes, 12328 cores
  - Dedicated 10-Gbps backbone network
  - 550 users and 120 publications per year

- A meta-cloud, meta-cluster, meta-data-center
  - Used by CS researchers in HPC, Clouds, Big Data, Networking, AI
  - To experiment in a fully controllable and observable environment
  - Similar problem space as Chameleon and Cloudlab (US)
  - Design goals
    - Support high-quality, reproducible experiments
    - On a large-scale, distributed, shared infrastructure
Landscape – cloud & experimentation

- Public cloud infrastructures (AWS, Azure, Google Cloud Platform, etc.)
  - 😞 No information/guarantees on placement, multi-tenancy, real performance

- Private clouds: Shared observable infrastructures
  - ☝ Monitoring & measurement
  - ✨ Ability to understand experiment results
  - 😞 No control over infrastructure settings

- Bare-metal as a service, fully reconfigurable infrastructure (Grid’5000)
  - ☝ Control/alter all layers (virtualization technology, OS, networking)
  - ✨ *In vitro* Cloud

And the same applies to all other environments (e.g. HPC)

---

1Inspired from a slide by Kate Keahey (Argonne National Lab)
Some recent results from Grid’5000 users

- Portable Online Prediction of Network Utilization (Inria Bdx + US)
- Energy proportionality on hybrid architectures (LIP/IRISA/Inria)
- Damaris: scalable, asynchronous data storage for large-scale simulations (Inria)
- Toward a resource management system for Fog/Edge infrastructures
Portable Online Prediction of Network Utilization

Problem
- Predict network utilization in near future to enable optimal utilization of spare bandwidth for low-priority asynchronous jobs co-located with an HPC application

Goals
- High accuracy, low compute overhead, learn on-the-fly without previous knowledge

Proposed solution
- Dynamic sequence-to-sequence recurrent neural networks that learn using a sliding window approach over recent history
- Evaluate the gain of a tree-based meta-data management
- INRIA, The Univ. of Tennessee, Exascale Comp. Proj., UC Irvine, Argonne Nat. Lab.

Grid’5000 experiments
- Monitor and predict network utilization for two HPC applications at small scale (30 nodes)
- Easy customization of environment for rapid prototyping and validation of ideas (in particular, custom MPI version with monitoring support)
- Impact: Early results facilitated by Grid’5000 are promising and motivate larger scale experiments on leadership class machines (Theta@Argonne)
Energy proportionality on hybrid architectures

- Hybrid computing architectures: low power processors, co-processors, GPUs...
- Supporting a "Big, Medium, Little" approach: the right processor at the right time

Damaris
Scalable, asynchronous data storage for large-scale simulations using the HDF5 format

▶ Traditional approach
  ♦ All simulation processes (10K+) write on disk at the same time synchronously
  ♦ Problems: 1) I/O jitter, 2) long I/O phase, 3) Blocked simulation during data writing

▶ Solution
  ♦ Aggregate data in dedicated cores using shared memory and write asynchronously

▶ Grid’5000 used as a testbed
  ♦ Access to many (1024) homogeneous cores
  ♦ Customizable environment and tools
  ♦ Repeat the experiments later with the same environment saved as an image
  ♦ The results show that Damaris can provide a jitter-free and wait-free data storage mechanism
  ♦ G5K helped prepare Damaris for deployment on top supercomputers (Titan, Pangea (Total), Jaguar, Kraken, etc.)
  ♦ https://project.inria.fr/damaris/
Toward a resource management system for Fog/Edge infras.

▶ Inria Project Lab: Discovery
♦ Design a resource management system (a.k.a. a cloudkit) for Fog/Edge infrastructures
♦ A four year project started in 2015 with Inria, Orange (and initially Renater)
♦ Designing from scratch such a system cannot be envisioned (OpenStack 13 Millions of LOCs)

▶ Contributions
♦ Implementation of a complete workflow to evaluate OpenStack WANWide scenarios
♦ Evaluate OpenStack up to 1000 compute nodes (Grid’5000, oct 2016)
♦ Evaluate OpenStack WANWide (impact of latency and throughput constraints) (oct 2017)
♦ Evaluation of communication bus for Fog/Edge scenarios (May 2018)
♦ Evaluation of database backends (NewSQL, NoSQL, etc. (May 2018)
Toward a resource management system for Fog/Edge infras.

**Inria Project Lab: Discovery (contd)**

- The creation of a dedicated working group within the OpenStack community that deals with Fog/Edge challenges (now managed by the foundation with key actors such as ATT, Verizon, CISCO, China mobile etc.)
- Several presentations / publications (see the DISCOVERY website)
- France has the main academic actor in the worldwide community (Inria/IMT Atlantique) thanks to the G5K testbed in particular.
  - A leadership position
  - A strong expertise for experiments related to performance, scalability of OpenStack components (concrete actions with RedHat, ongoing actions with Huawei, etc.)
An experiment’s outline

1. Discovering resources and selecting resources
2. Reconfiguring the resources to meet experimental needs
3. Monitoring experiments, extracting and analyzing data
4. Controlling experiments → automation, reproducible research
Discovering and selecting resources

- Describing resources ~ understand results
  - Covering nodes and network infrastructure
  - Machine-parsable format ~ scripts
  - Human-readable description on the web
  - Archived (State of testbed 6 months ago?)
  - Verified
    - Avoid inaccuracies/errors ~ wrong results
    - Self-checking by nodes before each reservation

- Selecting resources
  - Complex queries using resource manager
    oarsub -p "wattmeter='YES' and gpu='YES'"
    oarsub -l "\{cluster='a'\}/nodes=1+
                 \{cluster='b' and eth10g='Y'\}/nodes=2"

3https://www.grid5000.fr/w/Hardware
Reconfiguring resources

Operating System reconfiguration with Kadeploy:
- Provides a *Hardware-as-a-Service* cloud infrastructure
- Enable users to deploy their own software stack & get *root* access
- Scalable, efficient, reliable and flexible: 200 nodes deployed in ~5 minutes

Customize networking environment with KaVLAN
- Protect the testbed from experiments (Grid/Cloud middlewares)
- Avoid network pollution
- Create custom topologies
- By reconfiguring VLANS \(\leadsto\) almost no overhead

---

**KADEPLOY**

- Provides a *Hardware-as-a-Service* cloud infrastructure
- Enable users to deploy their own software stack & get *root* access
- Scalable, efficient, reliable and flexible: 200 nodes deployed in ~5 minutes

- Protect the testbed from experiments (Grid/Cloud middlewares)
- Avoid network pollution
- Create custom topologies
- By reconfiguring VLANS \(\leadsto\) almost no overhead

![Diagram](image)
Goal: enable users to understand what happens during their experiment

- System-level probes (usage of CPU, memory, disk, with Ganglia)
- Infrastructure-level probes: Kwapi
  - Power consumption, (Network traffic)
  - Captured at high frequency ($\approx 1$ Hz)
  - Live visualization
  - REST API
  - Long-term storage
Controlling experiments

- Legacy way of performing experiments: shell commands
  - 😞 time-consuming
  - 😞 error-prone
  - 😞 details tend to be forgotten over time

- Promising solution: **automation of experiments**
  - ~ Executable description of experiments
  - ~ Reproducible research

- Support from the testbed: Grid'5000 RESTful API *(Resource selection, reservation, deployment, monitoring)*

- Several higher-level tools to help automate experiments
  - Execo, Python-Grid5000 (Python), Ruby-cute (Ruby)

https://www.grid5000.fr/w/Grid5000:Software
Other features and services

- Flexible usage rules: default vs production (long jobs) vs besteffort (background)
- Data storage (group storage, disk reservation)
- Reproducible system images for Debian 8, 9, 10 (soon), testing; Centos 7, Ubuntu 18.04
- `sudo-g5k` to get root privileges on the standard environment
- To be announced soon:
  - Network-level federation with Fed4FIRE testbeds
  - 10 Gb connection to the Internet
- Work in progress:
  - New clusters in Nancy (3), Lyon, Grenoble
  - GPU-level reservations
  - IPv6
  - ARM servers
- 30 regression tests (1161 configurations) to ensure that everything works consistently
Key numbers: users and publications

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Active users</td>
<td>564</td>
<td>553</td>
<td>592</td>
<td>514</td>
<td>528</td>
<td>458</td>
<td>573</td>
<td>600</td>
<td>564</td>
</tr>
<tr>
<td>Publications</td>
<td>154</td>
<td>141</td>
<td>101</td>
<td>134</td>
<td>106</td>
<td>143</td>
<td>122</td>
<td>151</td>
<td>127</td>
</tr>
<tr>
<td>PhD &amp; HDR</td>
<td>14</td>
<td>20</td>
<td>9</td>
<td>27</td>
<td>24</td>
<td>30</td>
<td>27</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>Usage rate</td>
<td>50%</td>
<td>56%</td>
<td>58%</td>
<td>63%</td>
<td>63%</td>
<td>63%</td>
<td>55%</td>
<td>53%</td>
<td>70%</td>
</tr>
</tbody>
</table>

- 1313 active users over the last 3 years
- 3769 active users since 2003
- 2007 publications that benefited from Grid’5000 in our **HAL collection**
  - Computer Science: 96%, Mathematics: 2.4%, Physics: 2.4%

---

4https://hal.archives-ouvertes.fr/GRID5000
Grid’5000 organization and governance

Groupement d’Intérêt Scientifique (GIS)

- **Director** – Frédéric Desprez

- **Bureau** (6 members: FD, LN, Christian Perez, Adrien Lebre, Laurent Lefevre, David Margery)

- **Comité des responsables de sites**

- **Technical Director** – Lucas Nussbaum
  - Technical team (8.2 FTE)

- **Architects committee** (6 members)

- **Conseil de groupement**
  - Inria, CNRS, RENATER, CEA, CPU, CDEFI, IMT
  - (≈ Allistène + RENATER)

- **Conseil scientifique**
  - 10 international members

- institutional and scientific steering
- technical steering
- advisory and evaluation bodies
Technical organization

- Distributed infrastructure, but managed by a single distributed team
  - Strong coherence and coordination between sites
- Current composition: 8.2 full-time engineers
  - Inria: 5.91 (perm: 0.86, CDD: 5.1), CNRS: 1.02 (perm: 1.02), U. Rennes: 0.6 (perm: 0.6), IMT Atlantique: 0.4 (CDD: 0.4), U. Lorraine: 0.2 (perm: 0.2)
  - Including Pierre Neyron: Médaille de Cristal du CNRS 2019
Conclusions

- An advanced and established infrastructure for the *data-center* facets of Computer Science
  - Large-scale, distributed
  - Shared (many involved laboratories and institutions)
  - Designed for reconfigurability, observability, reproducible research

- Looking forward to extend this scope through collaborations in the context of SILECS
  - Compared with FIT: we share the same design goals, principles, but focus on different objects
    - *Core* of the Internet vs *Edge* of the Internet
    - *Internet of servers* vs *Internet of clients*
  - Strong needs of joint experiments