Adaptive and flexible platforms for e-Science - the future is cloudy

Johan Tordsson
Umeå University
tordsson@cs.umu.se

(joint work with Luis Tomás, Cristian Klein, Petter Svärd, Benoit Hudzia, Erik Elmroth, & Francisco Hernández)

October 16, 2014

www.cloudresearch.org
(eScience) challenge: Heterogeneity in computing

• Available platforms include
  – HPC Clusters
  – Grids
  – Accelerators (GPUs)
  – Clouds

• Programming models differ
  – Tools, languages, how to handle parallelisms/scalability, etc.

• Platforms also differ
  – Resource management systems
Resource management

- Batch system access
  - Waiting
- Dedicated resources
- Static
- Predictable?

- Instant access
- Shared resources
  - Noisy neighbors
- Dynamic?
- Unpredictable
Resource management

- Job completion time comparison

Image by Ian Foster
Addressing the challenge

Static -----> Dynamic
Unpredictable -----> Predictable

- More dynamic platforms?
- More adaptive applications?
- Collaboration?
  - Applications and platforms to interact
More dynamic platforms?

- **Hardware virtualization**
- **Virtual machines (VMs) & containers**
  - KVM, Xen, etc & Docker etc.
- **Application benefits**
  - Run multiple OS
    - Cross-platform execution
  - Supports different OS and library versions
  - Isolation
- **Platform benefits**
  - Consolidation, fewer servers
  - Simplified management, Isolation
  - Increased utilization, etc.
(Even) more dynamic platforms? - current hardware trends

- CPU, RAM, Disk access has “stopped” getting faster
- Network bandwidth is still increasing
- Remote resources get closer

Network Performance (Gbit/s)

- Source: http://www.intel.com/pressroom/kits/quickrefyr.htm
Server disaggregation

• Aggregate server resources into pools
  – Memory, Compute and I/O

• Easier provisioning and scaling
  – Scale by adding more resources from the pools
  – Free of single server limitations
Relaxed box model virtualization

For enhanced workload mixing (space)

Partial results

Modified KVM hypervisor
RDMA data transport
Infiniband network
Unmodified applications

Heca approach can provide resilience
8 GB VM, 33% remote

eScience?
- Big datasets?
- Contrast to parallel computing

Linux memory benchmark
SAP HANA, 2.5 TB (!)
In-memory DB
1 TB RAM vs 50% remote

Table 2: Overhead per query set with 80 HANA users.

<table>
<thead>
<tr>
<th>Demander VM</th>
<th>Sponsor VM</th>
<th>Sponsor VM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remote Memory</td>
<td>Remote Memory</td>
<td>Remote Memory</td>
</tr>
<tr>
<td>Qemu</td>
<td>Qemu</td>
<td>Qemu</td>
</tr>
<tr>
<td>Kernel</td>
<td>Kernel</td>
<td>Kernel</td>
</tr>
<tr>
<td>Heca Module</td>
<td>Heca Module</td>
<td>Heca Module</td>
</tr>
</tbody>
</table>

RDMA

Fig. 2: High-level architecture of a memory scale-out.

<table>
<thead>
<tr>
<th>Module</th>
<th>Hypervisor</th>
<th>Overhead</th>
<th>RAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qemu</td>
<td>Unmodified</td>
<td>4%</td>
<td>1G B:1G B:1G B</td>
</tr>
<tr>
<td>Heca</td>
<td>Modified KVM</td>
<td>0.9%</td>
<td>1G B:2G B:1G B</td>
</tr>
</tbody>
</table>

Error handling: if an application accesses an address that is not in

The kernel module handles the operations during scale-out and the transfer of memory content to and from remote hosts. The behavior of the kernel module differs based on whether it is running as a native program or as a virtualized program. When running as a native program, it directly interacts with the hardware memory management. When running as a virtualized program, it communicates with the hypervisor to manage memory operations. The hypervisor enables full transparency as it communicates cluster setup to the application in terms of a single address space, regardless of the physical memory distribution across the nodes.
More dynamic platforms?
- Overbooking

(Cloud) utilization issues

- Elasticity
- User overestimates
- T-shirt problem

![Graph showing utilization percentage and response time over time.](image)
Overbooking

- **Allocate more VMs than physical resources**
  - CPU, Memory, I/O
  - Higher VM consolidation
  - Increase resource utilization
Overbooking

- Admission control
  - How many VMs to accept?
- Scheduling
  - Where to allocate the VMs?

- Load predictions
- Fuzzy risk assessment
- Feedback to adjust (PID controllers)

More adaptive applications? Brownout - handle unexpected load variations

- 82% of users give up on lost payment transactions*
- 25% of end-users leave if load time > 4s**
- 20% reduced income if 0.5s longer load time***

- Impact on your eScience applications?

* JupiterResearch   ** Amazon   ***Google
Brownout - graceful performance degradation

- Disable optional functionality on a per-request level
- Goals:
  - Maximize optional computing
  - Avoid high response times
- eScience usage?
  - Mesh refinement?
  - Rate-adaptive computing?
  - Real-time simulations?

Application-Platform Collaboration?

BOB: Brownout + OverBooking

Collaboration - two approaches:
- naïve: brownout and overbooking used “as-is”
- smart: explicit communication is added
when the workload trend is predictable (Wikipedia-like) and sensitive, hence more prone to be disturbed by overbooking inside an 8-cores VM. The other 16 cores host jobs. Notably, assigned to services, i.e., RUBiS and RUBBoS each deployed Fig. 2 (a) and (b). This creates different trends and peaks, and run for a long time (months, years) and are accessed by

---

**Fig. 2 (c) Naïve BOB**

L Tomás, C Klein, J Tordsson, F Hernandez, The straw that broke the camel’s back: safe cloud overbooking, *Cloud and Autonomic Computing (CAC) 2014*
Conclusions & perspectives

• Future computing platform will be diverse and more dynamic
• New (programming) paradigms will appear

• Adaptive applications key
• Applications must collaborate with platforms
  – Careful design of interacting adaptive systems required