The Grand CRU Challenge

Marcel Blöcher\textsuperscript{1}, Malte Viering\textsuperscript{1}, Stefan Schmid\textsuperscript{2}, and Patrick Eugster\textsuperscript{1\&3}

\textsuperscript{1}TU Darmstadt, Germany
\textsuperscript{2}Aalborg University, Denmark & TU Berlin, Germany
\textsuperscript{3}Purdue University, USA
Cluster Resource Scheduling

Scheduling information is distributed!

demands

resources

placement

Scheduler A

Scheduler B

job\textsubscript{A}

job\textsubscript{B}

job\textsubscript{C}

e.g. Clos

cluster state

local resources

network resources
Why bother?

Two-dimensional resource scaling: an Apache Kafka streaming case study
Cluster Resource Utilization

What is required for taking informed resource scheduling decisions?

CRU dilemma
Without knowledge of both roles’ information, scheduling decisions are likely to be suboptimal.

But both options speak against a clear separation

Application Information
- performance goals
- resources usage

Cluster Information
- node-local resources
- network resources

enrich the application

applications learn more about the underlying infrastructure → schedule an entire “graph” of containers

enrich the resource manager

resource manager understand more of the applications’ semantics and performance goals
The Grand CRU Challenge

*Idea*: Share slightly more information but

- respect separation of different roles
- naïve approach (expose all information) becomes combinatorial and expensive
- resources are different in nature - shared vs local resources

**Challenge**: Find a cluster scheduling architecture which provides efficient information sharing mechanisms
Multi-Dimensional Scheduling

Local resources can be handled in an isolated fashion. Network resources are shared and allocation is intertwined with that of local resources.

What are the consequences for the scheduler architecture?
Design Space: Scheduling Architectures

Monolithic

Two-Level

Shared-State
CRU Dilemma - Evaluation
Two-Level Architecture

1. Resource Hoarding Issue

2. Resource Offer Conflict

- network shares in an offer affect more than a single node

next slide
Resource Offer Conflict

T1 a job runs 6 tasks
T2 blue offer, spawn new task

T3 in the meantime, a tasks finishes
T4 response to offer

Bandwidth demand
differs, i.e. response to offer affects other resources

CRU Dilemma - Evaluation
Shared-State Architecture

- Simulation based evaluation
  - modified Omega simulator, network perspective added
  - each job’s task $\rightarrow$ VC bandwidth demand
  - 6000 node Fat-Tree, avg. 200 tasks per job
  - 2 schedulers running simultaneously

- Metrics
  - scheduler busy time
  - conflict fraction of scheduling decisions
Shared-State Scheduler

Experiment A: only **node-local** resources
Shared-State Scheduler

Experiment B: network+local resources

scheduling time: production system estimates reported in the Omega paper
Conclusion

- We make the case for multi-dimensional resource scheduling
- Scheduling information is distributed → CRU dilemma

**Open Question - Grand CRU Challenge:**
*How to maximize CRU when networking enters the picture?*

<table>
<thead>
<tr>
<th>Issue</th>
<th>Core Design Principle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolithic</td>
<td>single point which holds all information</td>
</tr>
<tr>
<td>does not scale / multi-path issue</td>
<td></td>
</tr>
<tr>
<td>Two-Level</td>
<td>distributed, by small disjoint information shares</td>
</tr>
<tr>
<td>too pessimistic</td>
<td></td>
</tr>
<tr>
<td>Shared-State</td>
<td>distributed write access by conflict resolution</td>
</tr>
<tr>
<td>too many conflicts</td>
<td></td>
</tr>
</tbody>
</table>

None of the investigated architectures tackles the CRU dilemma

We advocate an architecture that combines all three design principle
Thank You

contact: bloecher@dsp.tu-darmstadt.de

Marcel Blöcher\textsuperscript{1}, Malte Viering\textsuperscript{1}, Stefan Schmid\textsuperscript{2}, and Patrick Eugster\textsuperscript{1\&3}

\textsuperscript{1}TU Darmstadt, Germany
\textsuperscript{2}Aalborg University, Denmark & TU Berlin, Germany
\textsuperscript{3}Purdue University, USA