Scheduling the I/O of HPC applications under congestion

Ana Gainaru, Guillaume Aupy, Anne Benoit, Yves Robert, Franck Cappello & Marc Snir

Motivation

Model
Platform
Applications

Objectives

Algorithms

Cimulation

Applications

Assessment of heuristics

Experiments

Conclusion

Motivation

2 Model Platform Applications

- 6 Algorithms
- 4 Simulations
 Applications
 Assessment of heuristics
- 5 Experiments
- 6 Conclusion

Interconnect technologies: A major challenge

G. Aupy

Motivation

Platform

Applications Objectives

Algorithm

Simulation

Applications

Assessment of heuristics

Experiment

Conclusion

Application interference - Performance degradation due to applications accessing a shared resource (PFS)

Some app observe up to 14x decrease of performance due to interference

The challenge:

Flops are "free", we need to optimize data-movement!

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

C: 1 ...

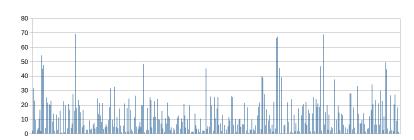
Applications
Assessment of

heuristics

Experiment

Conclusion

Interconnect technologies: A major challenge



Analysis of the Intrepid system @Argonne: I/O throughput decrease (percentage per application, over 400 applications).

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

Simulation

Applications
Assessment of
heuristics

Experiments

Conclusion

1 Motivation

- 2 Model
 Platform
 - Applications
 Objectives
- Algorithms
- 4 Simulations
 Applications
 Assessment of heuristics
- 6 Experiments
- **6** Conclusion

в

Motivation

Model Platform

Applications Objectives

Algorithms

Cimandakian

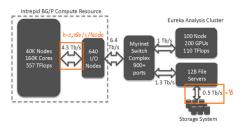
Applications Assessment of

heuristics

Experiment

Conclusion

- N unit-speed processors, equipped with an I/O card of bandwidth b
- Centralized I/O system with total bandwidth B



Model instantiation for the Intrepid platform.

Mode

Applications

Algorithn

Simulation

Applications

-

Experiment

Conclusion

K applications competing for I/O. For application $App^{(k)}$:

- Released at time r_k ;
- Executed on $\beta^{(k)}$ procs;
- $n_{\text{tot}}^{(k)}$ instances: $\mathcal{I}_i^{(k)}$ consists of $w^{(k,i)}$ units of computation followed by the transfer of a volume $\text{vol}_{io}^{(k,i)}$;
- The minimum time to execute $vol_{in}^{(k,i)}$ is:

$$\mathsf{time}_{\mathsf{io}}^{(k,i)} = \frac{\mathsf{vol}_{\mathsf{io}}^{(k,i)}}{\mathsf{min}(\beta^{(k)}b, B)};$$

• Last instance finishes at time d_k .



Applications

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

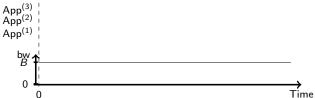
Simulations

Applications Assessment of heuristics

G

Experiments

Conclusion



Applications

G. Aupy

Motivation

Model

Platform
Applications
Objectives

Algorithms

Simulations

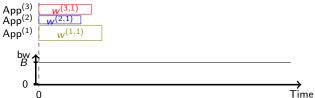
Applications

Applications
Assessment of
heuristics
Appl

G

Experiments

Conclusion



0

. . .

Applications

G. Aupy

Motivation

Model

Platform
Applications
Objectives

Algorithms

Simulations

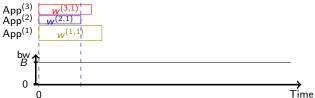
Applications
Assessment of
heuristics
App

G

- Incurrence

Experiments

Conclusion



Motivation

Model

Platform Applications

Objectives

Algorithms

Simulations

Applications

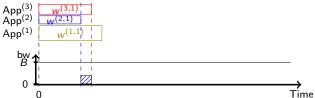
Assessment of heuristics

G

Experiments

Experiment

Conclusion



Motivation

Model

Platform
Applications
Objectives

Algorithms

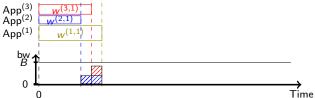
Simulations

Jiiiulation

Applications
Assessment of heuristics
App

Experiments

Conclusion



1/0 scheduling **Applications**

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

Simulations

Applications

G

Assessment of heuristics

Experiments

Conclusion



Motivation

Model

Platform Applications

Objectives

Algorithms

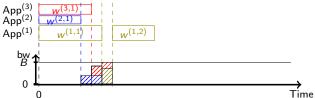
Simulations

Applications
Assessment of
heuristics
App

G

Experiments

Conclusion



Motivation

Model

Platform Applications

Objectives

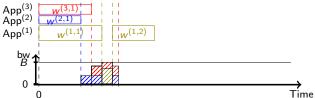
Algorithms

Simulations

Applications
Assessment of
heuristics
Appl

Experiments

Conclusion



Motivation

Model

Platform
Applications
Objectives

Algorithms

Simulations

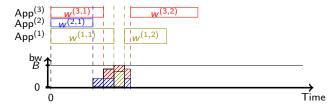
Applications

Assessment of heuristics

Experiments

Conclusion

G



Motivation

Model

Platform Applications

Objectives

Algorithms

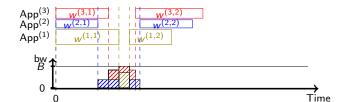
Simulations

Applications

Assessment of heuristics

Experiments

Conclusion



Motivation

Model

Platform
Applications
Objectives

Algorithms

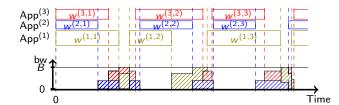
Simulations

Applications

Assessment of heuristics

Experiments

Conclusion



Model

Platform Applications

Objectives

Algorithms

Cimulation

Applications Assessment of

F.....

Conclusion

Definition (Application efficiency)

$$\tilde{\rho}^{(k)}(t) = \frac{\sum_{i \leq n^{(k)}(t)} w^{(k,i)}}{t - r_k},$$

where $n^{(k)}(t)$ is the number of instances of $App^{(k)}$ executed at time t.

Model

Platform Applications Objectives

Algorithn

Cimulation

Applications Assessment of

Evperiment

Conclusion

Definition (Application efficiency)

$$\tilde{\rho}^{(k)}(t) = \frac{\sum_{i \leq n^{(k)}(t)} w^{(k,i)}}{t - r_k},$$

where $n^{(k)}(t)$ is the number of instances of $App^{(k)}$ executed at time t.

Obviously: $t - r_k \ge \sum_{i \le n^{(k)}(t)} \left(w^{(k,i)} + \mathsf{time}_{\mathsf{io}}^{(k,i)} \right)$.

Hence:

$$\tilde{\rho}^{(k)}(t) \leq \rho^{(k)}(t) = \frac{\sum_{i \leq n^{(k)}(t)} w^{(k,i)}}{\sum_{i \leq n^{(k)}(t)} \left(w^{(k,i)} + \mathsf{time}_{\mathsf{io}}^{(k,i)} \right)}.$$

Model Platform

Applications
Objectives

Algorithms

Simulation

Applications
Assessment of

Experiment

Conclusion

• SysEfficiency:

maximize
$$\frac{1}{N} \sum_{k=1}^{K} \beta^{(k)} \tilde{\rho}^{(k)}(d_k)$$
.

• DILATION:

$$\text{minimize } \max_{k=1..K} \frac{\rho^{(k)}(d_k)}{\tilde{\rho}^{(k)}(d_k)}.$$

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

Simulation

Applications Assessment of

heuristics

Experiments

Conclusion

- Motivation
- 2 Model Platform Applications Objectives
- 3 Algorithms
- 4 Simulations
 Applications
 Assessment of heuristics

G

- 6 Experiments
- 6 Conclusion

Mode

Platform Applications Objectives

Algorithms

Simulation

Applications

F....

Conclusion

The scheduler monitors the stream of I/O calls; decides on the fly which applications can perform I/O.

- At each time step, it has access to the state of the system (each application efficiency, $\tilde{\rho}^{(k)}$).
- Based on a given strategy, chooses a subset of applications that are allowed to perform I/O.

Applications Objectives

Algorithms

Applications

The scheduler monitors the stream of I/O calls; decides on the fly which applications can perform I/O.

- At each time step, it has access to the state of the system (each application efficiency, $\tilde{\rho}^{(k)}$).
- Based on a given strategy, chooses a subset of applications that are allowed to perform I/O.

When a strategy favors $App^{(k)}$, it means that $App^{(k)}$ is executed as fast as possible (min $(b\beta^{(k)}, bw_{avail})$).

Motivation

Model

Platform Applications Objectives

Algorithms

C

Applications Assessment of

heuristics

Conclusion

 ROUNDROBIN: Similar to the current scheduler in HPC systems. Applications are served following the "First-Come, First Served" principle.

8 0

Motivation

Model

Applications Objectives

Algorithms

Simulation

Applications

Conclusion

- ROUNDROBIN: Similar to the current scheduler in HPC systems. Applications are served following the "First-Come, First Served" principle.
- MINDILATION: favors applications with high values of $\frac{\rho^{(k)}(t)}{\tilde{\sigma}^{(k)}(t)}$.

8 0

Model

Applications Objectives

Algorithms

Simulation

Applications

Conclusion

- ROUNDROBIN: Similar to the current scheduler in HPC systems. Applications are served following the "First-Come, First Served" principle.
- MINDILATION: favors applications with high values of $\frac{\rho^{(k)}(t)}{\tilde{\rho}^{(k)}(t)}$.
- MAXSYSEFF: favors applications with low values of $\beta^{(k)}\tilde{\rho}^{(k)}(t)$.

Mode

Platform Applications Objectives

Algorithms

Simulation

Applications
Assessment of

F. ...

Conclusion

- ROUNDROBIN: Similar to the current scheduler in HPC systems. Applications are served following the "First-Come, First Served" principle.
- MINDILATION: favors applications with high values of $\frac{\rho^{(k)}(t)}{\tilde{\rho}^{(k)}(t)}$.
- MAXSYSEFF: favors applications with low values of $\beta^{(k)}\tilde{\rho}^{(k)}(t)$.
- MINMAX- γ : same as MAXSYSEFF, unless there exists an applications with $\frac{\tilde{\rho}^{(k)}(t)}{\rho^{(k)}(t)}$ below a threshold γ . In that case, switches to MINDILATION.

3 0 0 0 0 0 0 0

Mode

Platform Applications Objectives

Algorithms

Simulation

Applications

Evperiment

Conclusion

- ROUNDROBIN: Similar to the current scheduler in HPC systems. Applications are served following the "First-Come, First Served" principle.
- MINDILATION: favors applications with high values of $\frac{\rho^{(k)}(t)}{\tilde{\rho}^{(k)}(t)}$.
- MAXSYSEFF: favors applications with low values of $\beta^{(k)}\tilde{\rho}^{(k)}(t)$.
- MINMAX- γ : same as MAXSYSEFF, unless there exists an applications with $\frac{\tilde{\rho}^{(k)}(t)}{\rho^{(k)}(t)}$ below a threshold γ . In that case, switches to MINDILATION.

PRIORITY variant: if an application has started to do some I/O, then it is prioritized.

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

Simulations Applications

Assessment of heuristics

Experiments

Experiment

Conclusion

- Motivation
- 2 Model Platform Applications Objectives
- 3 Algorithms
- 4 Simulations
 Applications
 Assessment of heuristics

G

- 5 Experiments
- 6 Conclusion

Motivation

Model

Platform Applications Objectives

Algorithms

C1............

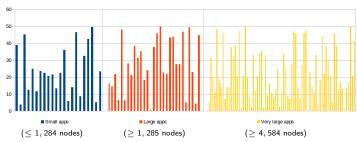
Applications Assessment of

heuristics

Experiments

Conclusion

We use Darshan to capture the behavior of applications that ran on Intrepid (2013).



Percentage time spent doing I/O per application type.

Motivation

Model

Platform Applications Objectives

Algorithms

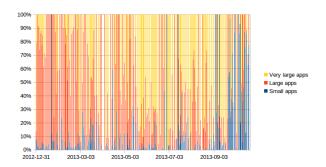
C:....lations

Applications
Assessment of

Evneriment

Conclusion

We use Darshan to capture the behavior of applications that ran on Intrepid (2013).



System usage per day for each application type

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

Simulations

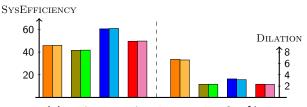
Applications
Assessment of
heuristics

Experiment

Conclusion







(a) 10 large applications, ratio of 20%Objectives for different mixes of applications and I/O computation ratios.

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

Simulations

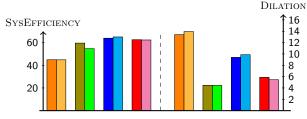
Applications
Assessment of
heuristics

Experimen

Conclusion







(b) 50 small and 5 large applications, ratio of 20%

Objectives for different mixes of applications and I/O computation ratios.

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

Simulations

Applications

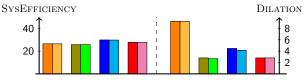
Assessment of heuristics

Experimer

Conclusion







(c) 50 small and 5 large applications, ratio of 35%Objectives for different mixes of applications and I/O computation ratios.

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

Simulation

Applications
Assessment of

Assessment o heuristics

Conclusion

Comparison of the heuristics on current platforms

We then compared our results with the Intrepid and Mira scheduler when congestion occurs.

Note that Intrepid and Mira use an architectural enhancement to improve the behavior of applications with large bursts of I/O: *Burst Buffers*.

1/0 scheduling

Comparison of the heuristics on current platforms

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

Applications Assessment of heuristics

Conclusion

	DILATION	SysEfficiency
	(minimize)	(maximize)
MaxSysEff	2.46	85.35
PRIORITY variant	3.13	82.98
MinMax-0.25	2.33	83.08
PRIORITY variant	2.93	80.31
MinMax-0.5	1.99	77.2
PRIORITY variant	2.43	72.96
MinMax-0.75	1.69	71.66
PRIORITY variant	2.03	66.94
MinDilation	1.63	70.45
PRIORITY variant	1.96	65.64
Intrepid	2.55	71.12
Upper-limit	_	91.59

Table: Averages over 56 different congested moments on Intrepid.

G. Aupy

Motivation

Model

Platform Applications Objectives

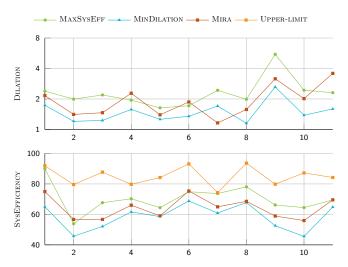
Algorithms

Simulations

Applications
Assessment of heuristics

Experimen

Conclusion



Comparison of the PRIORITY heuristics over the current DILATION and SYSEFFICIENCY of Mira.

. 6

G. Aupy

Motivation

Model Platform

Platform Applications Objectives

Algorithms

Simulation

Applications Assessment of

heuristics

Experiments

Experiment

Conclusion

- Motivation
- 2 Model Platform Applications Objectives
- Algorithms
- 4 Simulations Applications Assessment of heuristics
- **5** Experiments
- 6 Conclusion

G. Aupy

Motivation

Model Platform

Applications Objectives

Algorithms

Simulation

Applications Assessment of

heuristics

Experiments

Conclusion

- Experiments on Vesta (development platform for Mira)
- ullet Vesta is using hard disks and is affected by locality: we only used the $\ensuremath{\mathrm{PRIORITY}}$ variant of heuristics
- \bullet We implemented the heuristics as an additional layer on top of Vesta I/O scheduler

. . 6

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

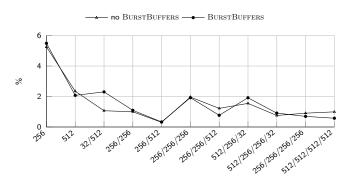
Cimandakian

Applications Assessment of

Experiments

Conclusion

- Experiments on Vesta (development platform for Mira)
- \bullet Vesta is using hard disks and is affected by locality: we only used the Priority variant of heuristics
- We implemented the heuristics as an additional layer on top of Vesta I/O scheduler



Execution time overhead of our implementation of the IOR benchmark.

G. Aupy

Motivation

Model

Platform Applications Objectives

Algorithms

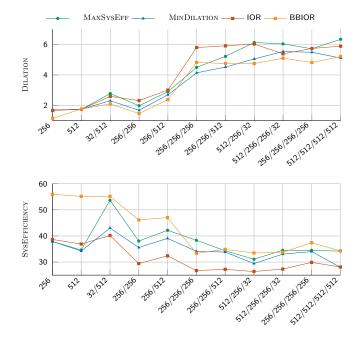
Simulation

Applications

Assessment of heuristics

Experiments

Conclusion



System efficiency and dilation for different scenarios on Vesta.

G. Aupy

Motivation

Model

Platform Applications Objectives

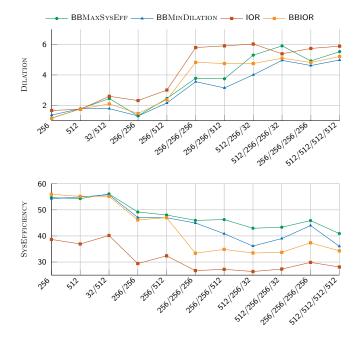
Algorithms

Cimulation

Applications
Assessment of

Experiments

Conclusion



System efficiency and dilation for different scenarios on Vesta.

G. Aupy

Motivation

Model Platform

Applications Objectives

Algorithms

Simulation

Applications
Assessment of
heuristics

Experiments

Experiment

Conclusion

- Motivation
- 2 Model Platform Applications Objectives
- Algorithms
- 4 Simulations
 Applications
 Assessment of heuristics
- 6 Experiments
- **6** Conclusion

Model

Platform Applications Objectives

Algorithms

Simulation

Applications
Assessment of

heuristics

Conclusion

- New I/O scheduler taking global view of system into account
- Outperforms current scheduler
- More experiments needed on larger application sets
- Window-based schedules for periodic applications?