The XtreemOS JScheduler: Using Self-Scheduling Techniques in Large Computing Architectures

Barcelona Supercomputing Center
Technical University of Catalunya

F. Guim, I. Rodero, M. Garcia, J. Corbalan

Outline

- The XtreemOS Project
- The scenario and its challenges
- The ISIS-Dispatcher
- Including the ISIS-Dispatcher in the XOS
- Evaluation
- Conclusions & Future Work

Outline

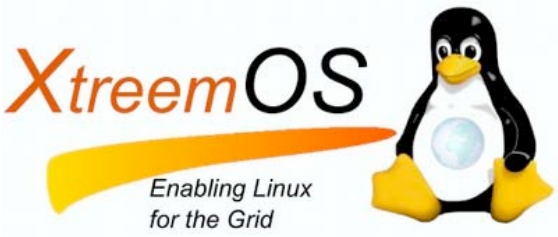
Francesc Guim Bernat

- The XtreemOS Project
- The scenario and its challenges
- The ISIS-Dispatcher
- Including the ISIS-Dispatcher in the XOS
- Evaluation
- Conclusions & Future Work

The project

Francesc Guim Bernat

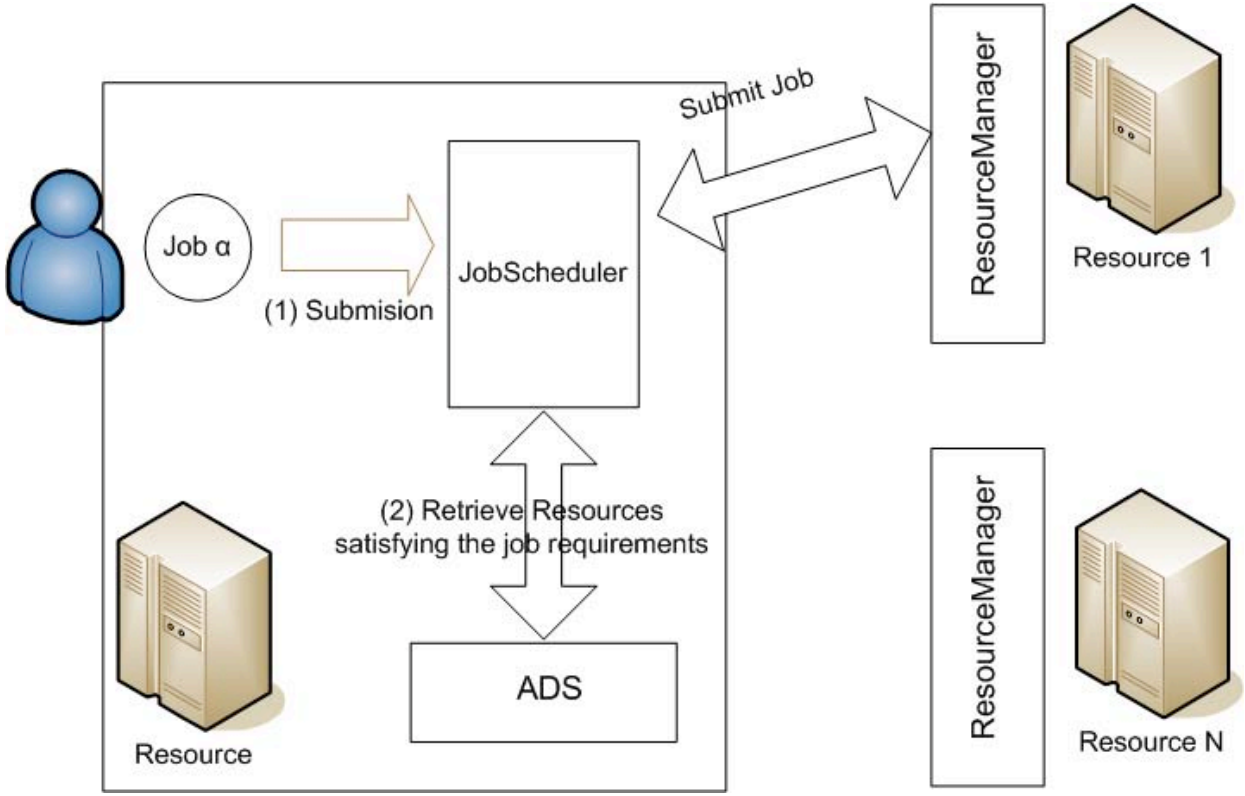
- It has been already introduced …
- It aims at investigating and proposing new services that should be added to current operating systems to build large Grid.
- In this paper we focus on:
  - the Application Execution Management (AEM) component of the XOS responsible of:
    - Job scheduling
    - Resource management.
  - Job Scheduling Strategies for this system
    - How we deal with job submissions in such large systems ?



The AEM Architecture

Francesc Guim Bernat

- ADS → Application Discovering System
- jScheduler → schedules one job, it receives a pre-selection of resources from the ADS
- Resource Manager → manages the computational resource

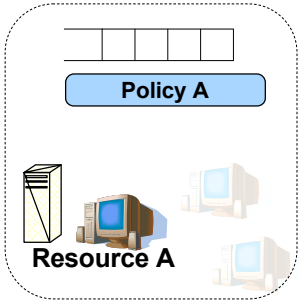


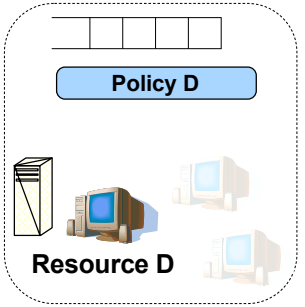
Outline

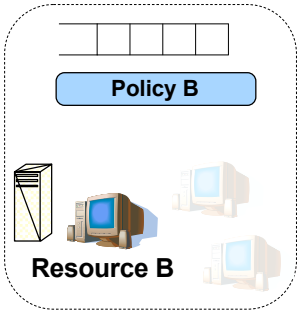
- The XtreemOS Project
- The scenario and its challenges
- The ISIS-Dispatcher
- Including the ISIS-Dispatcher in the XOS
- Evaluation
- Conclusions & Future Work

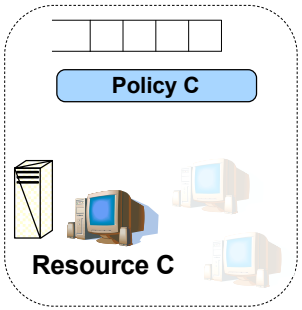
The Architecture

- *N* Independent centers
  - No centralized architectures
  - keep their scheduling policies
  - heterogeneous with different capabilities
  - Submission: Local Centers or Dispatcher
- The scheduling has to deal with
  - Large scale systems
  - Dynamic systems
  - Very Heterogeneous

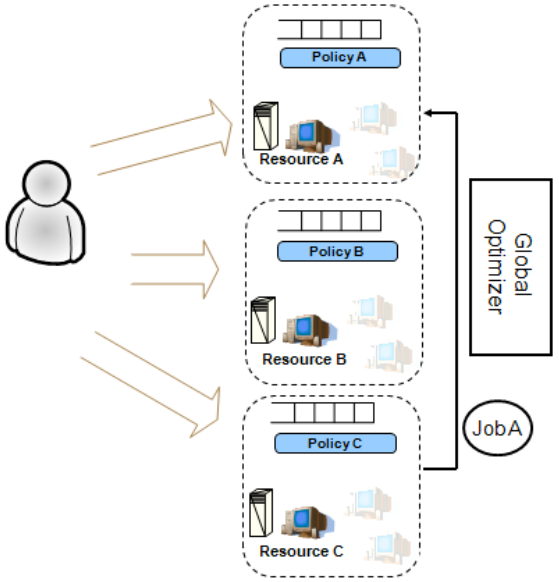




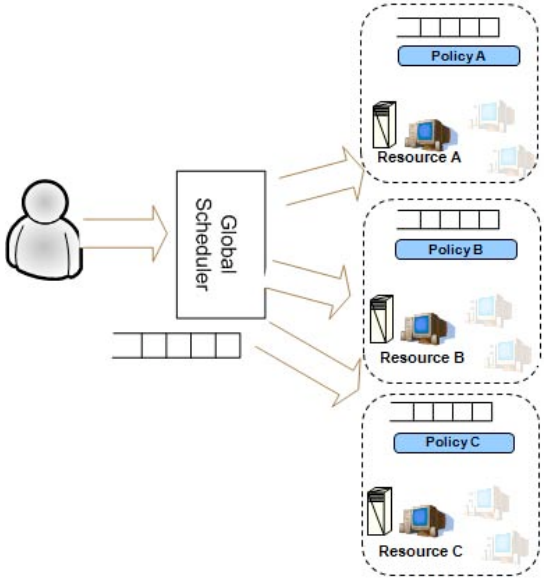




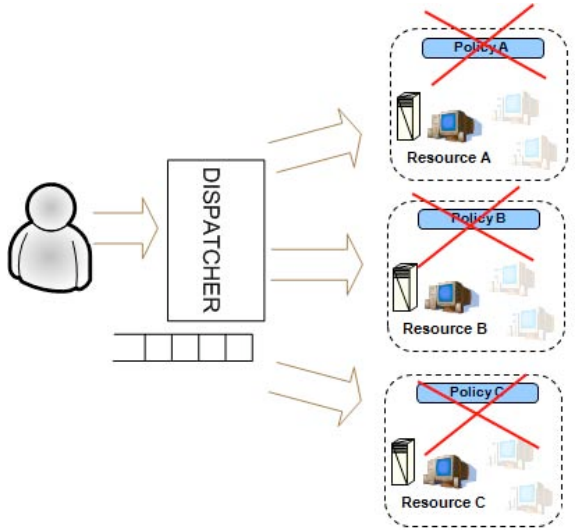
Proposed Solutions



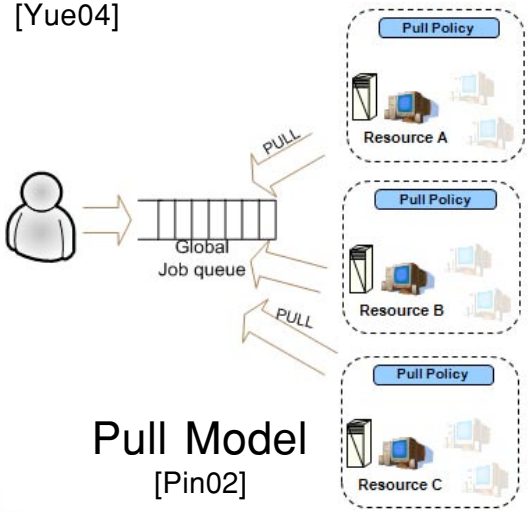
Global Optimizer Model
[Yue04]



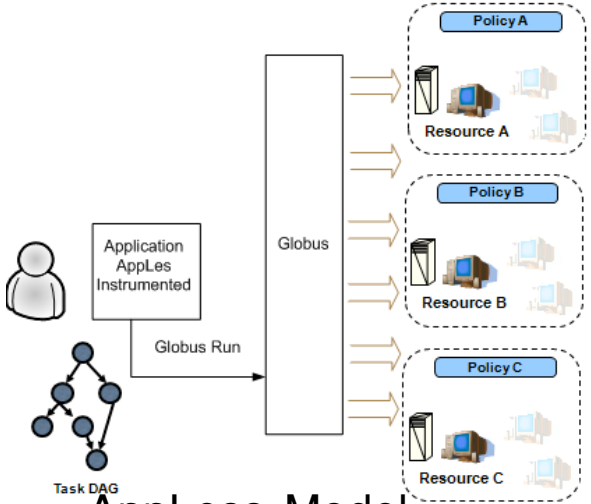
Global Scheduler Model
[Rod05,Huang97,Diet01]



Global Dispatcher Model
[Car04,Shro04]



Pull Model
[Pin02]



AppLess Model
[Ber07]

Outline

- The XtreemOS Project
- The scenario and its challenges
- The ISIS-Dispatcher
- Including the ISIS-Dispatcher in the XOS
- Evaluation
- Conclusions & Future Work

Our proposal

Francesc Guim Bernat

- The ISIS-Architecture
  - Optimize user metrics
  - One Dispatcher per job
  - Task Dispatching Policies
  - Local Scheduling information → New API between Dispatcher/HPC Centers
  - Use of Advanced Services (i.e: Runtime predictors)
- User metrics to optimize
  - Wait time
  - Slowdown
  - Etc.

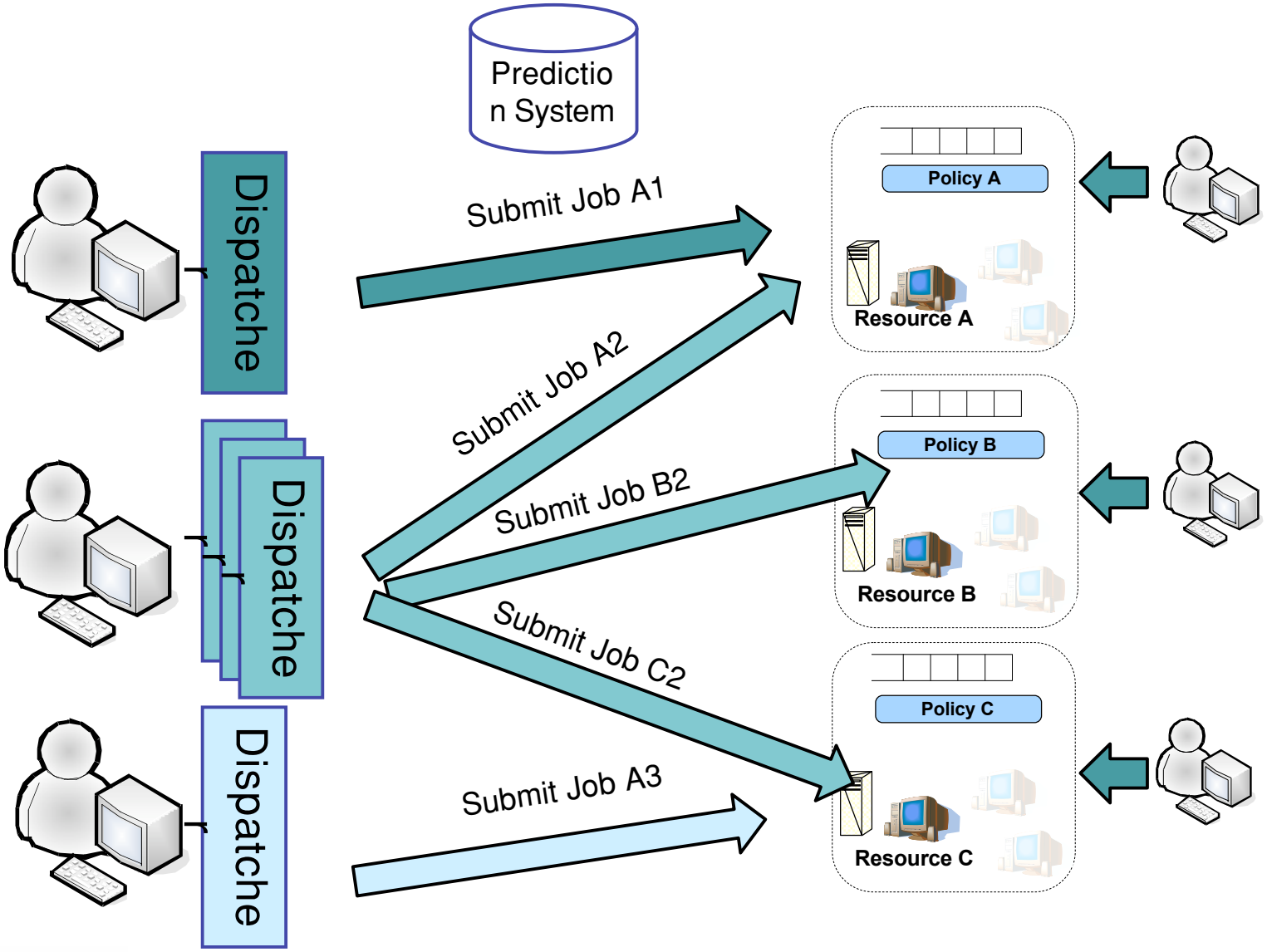
Task Dispatching Policies

- Random [mark99, harch00,aguilar97]
- Round-Robin [mark99, harch00,mark98]
- Shorts-Queue [schro00,harch99]
- Less Work Left [schro00,harch99]
- Less Submitted Jobs [schro00,harch99]
- ...

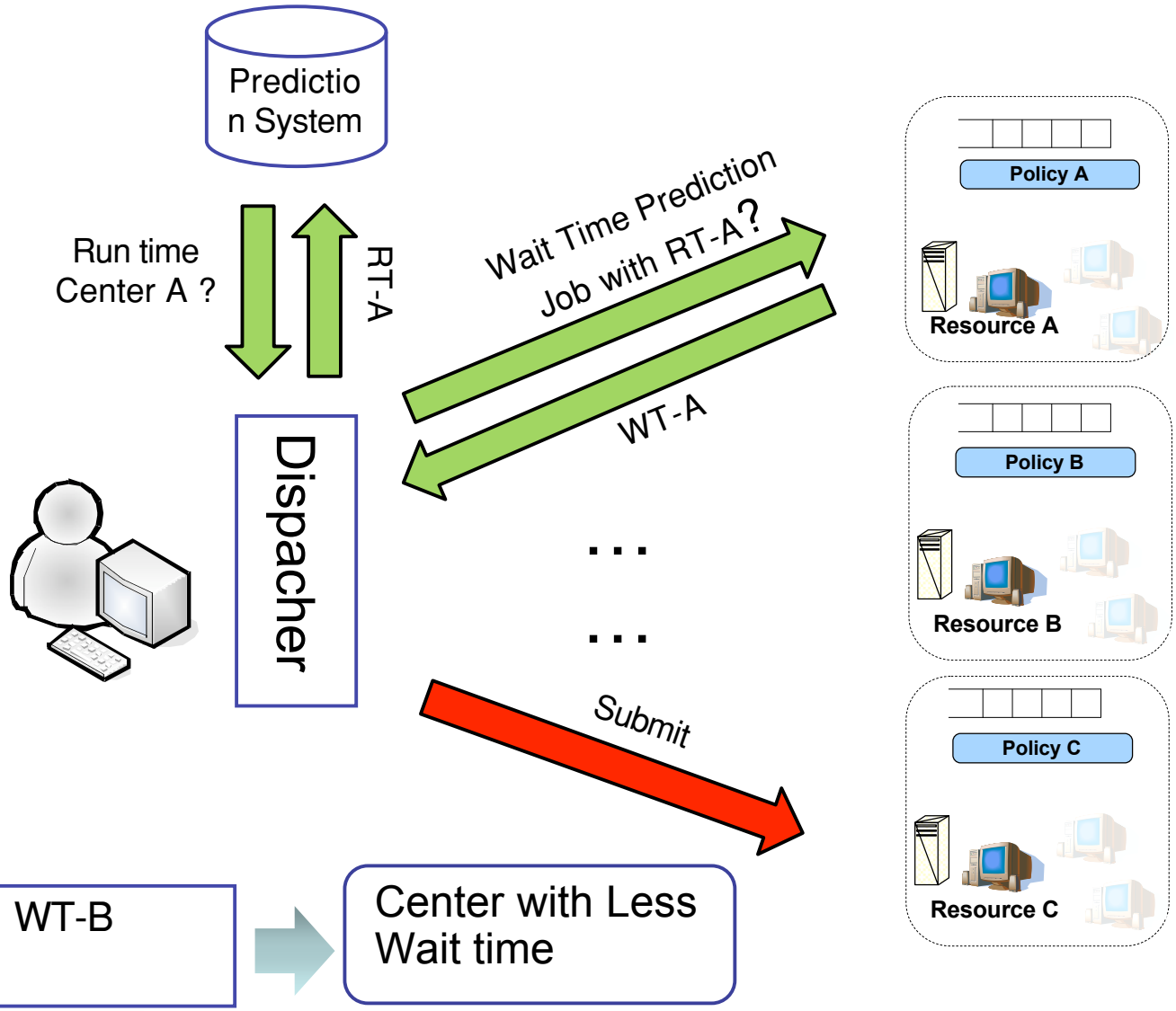
All based on the System Status Information

We propose to use Scheduling Information

The ISIS Dispatcher techniques



Scheduling Based on the Wait time



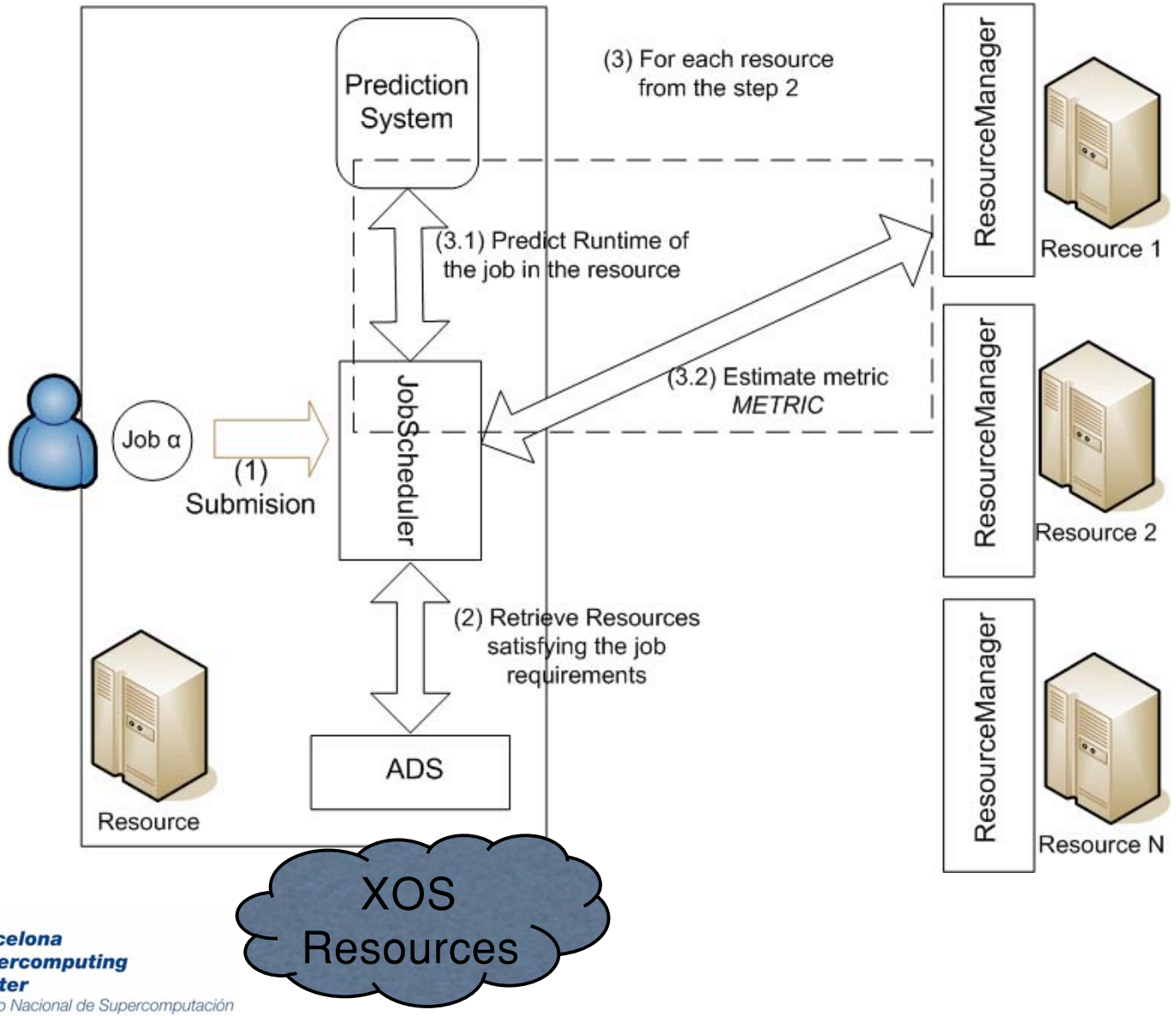
WT-A WT-B WT-C

Center with Less Wait time

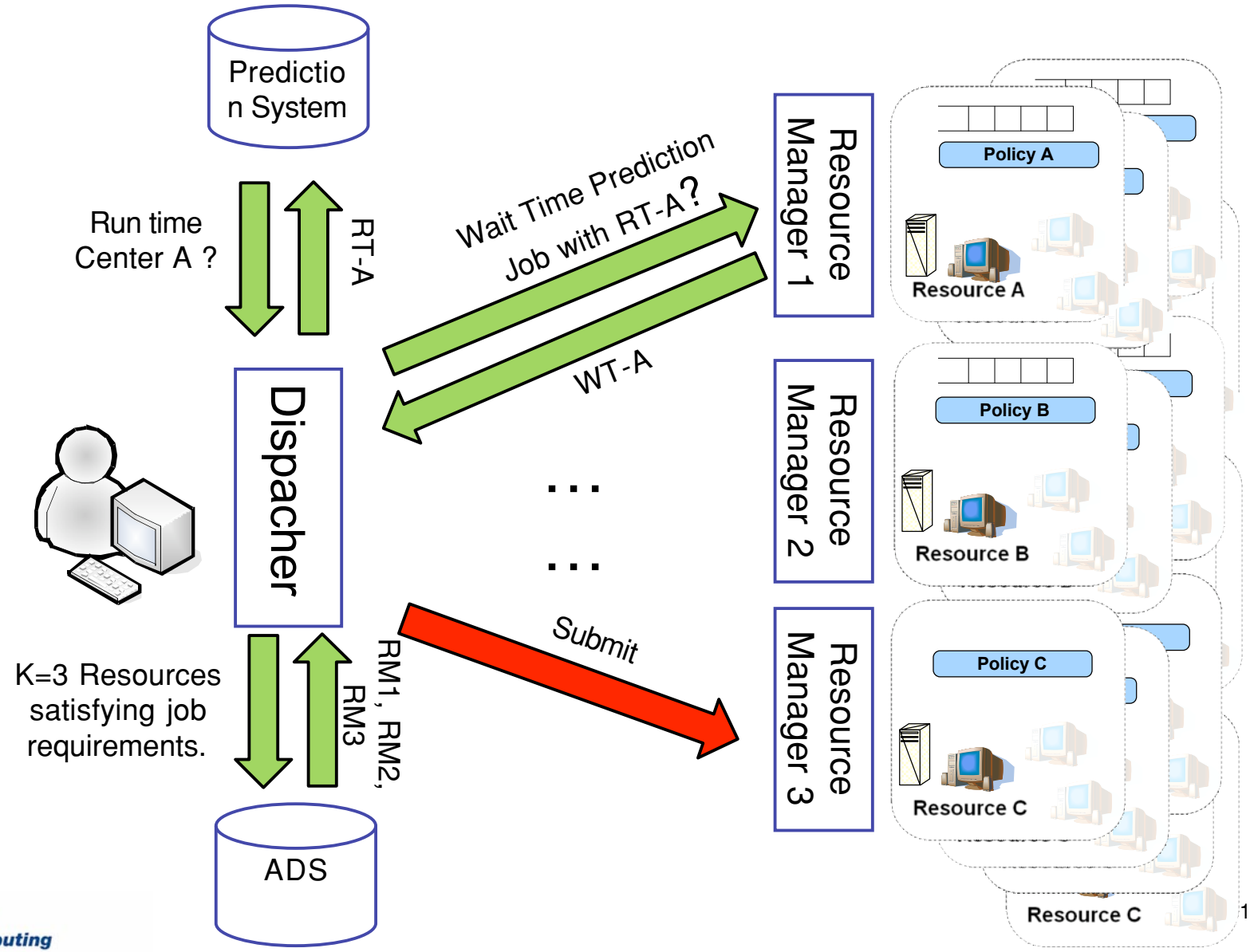
Outline

- The XtreemOS Project
- The scenario and its challenges
- The ISIS-Dispatcher
- Including the ISIS-Dispatcher in the XOS
- Evaluation
- Conclusions & Future Work

The Xtreem OS Extension



The new Scheduling



Outline

- The XtreemOS Project
- The scenario and its challenges
- The ISIS-Dispatcher
- Including the ISIS-Dispatcher in the XOS
- Evaluation
- Conclusions & Future Work

The evaluation model

- Alvio Simulator
  - Event Driven Simulator
  - Models all the components of multi-sites systems
    - F. Guim, J. Corbalan JSSPP 07
    - F. Guim, I.Rodero Grid 08
  - Models the local resources (Local Resource Managers + Schedulers)
    - F. Guim, J. Corbalan, J. Labarta, PDCAT 2007
    - F. Guim, J. Corbalan. HPCS 08
- Workloads used
  - Cluster & Grid architectures
  - Standard Workload Format [Steve99]
  - Workload Archive [www.cs.huji.ac.il/~feit/parallel/workload/]

In the model we have ..

- K independent centers:
  - Number of processors
  - Performance Factor
  - Job Scheduling Policy (FCFS, SBF-Backfilling, EASY-Backfilling, Shortest Job First and LXWF-Backfilling)
  - Resource Selection Policy (First Fit)
- The prediction system
  - Uses classification trees + discretization techniques
- We have modeled the ADS
  - Interface

    ListOfRM resourcesMatching(JobRequirements, int *k*);
  - The ADS returns K Resource Managers

    SelectedRM~U[1..N] → N number of centers

The Workloads & Scenarios

- The NASA Ames iPSC/860 log.
- The Los Alamos National Lab (LANL-CM5) log.
- The San-Diego Supercomputer Center Paragon (SDSC-Par).
- The Cornell Theory Center (CTC) SP2 log.
- The Lawrence Livermore National Lab (LLNL).
- The Swedish Royal Institute of Technology (KTH) IBM SP2 log.
- The San Diego Supercomputer Center (SDSC-SP2) SP2 log.
- The LANL Origin 2000 Cluster (Nirvana) log.
- The OSC Linux Cluster log (OSC).
- The San Diego Supercomputer Center Blue Horizon log
- The HPC2N log.
- The DAS2 5-Cluster Grid Logs.
- The San Diego Supercomputer Center DataStar log
- The LPC Log.
- The LCG Grid log.
- The SHARCNET log .
- The LLNL Atlas log.
- The LLNL Thunder log.

Center	CPUs	Fact.	Policy
NASA Ames	128	4	SJBF
LANL-CM5	1024	1	FCFS
SDSC Paragon	416	1	EASY
CTC IBM SP2	512	2	EASY
KTH	100	4	EASY
SDSC SP2	128	4	CNWF
Nirvana	2048	4	EASY
OSC	178	4	SJBF
SDSC-Blue	1024	2	FCFS
HPC2N	240	4	EASY
DAS-fs0	144	4	EASY
DAS-fs1	64	1	SJF
DAS-fs2	64	5	SJBF
DAS-fs3	64	1	SJBF
DAS-fs4	64	5	FCFS
SDSC-DS	184	5	
LPC	70 x 2	8	FCFS
LGC	100 x 250	5	EASY
Sharnet	6 x 128 1 x 1068 1 x 1536 1 x 3072 1 x 384	5 5 5 5 5	SJF SJF SJF SJF SJF
Atlas	1152	5	FCFS
Thunder	1024	8	EASY
CM5	1152	3	FCFS

The Experiments

- Evaluation of the Local Centers (original scenarios)
- Evaluation of the XOS+ISIS Architecture
  - The ADS
    - Returning K Resource Managers Selected Randomly
  - One dispatcher per job
  - One prediction system
- The Task dispatching policies
  - Less-Waittime
    - Based on runtime prediction
  - Less-Slowdown
    - Based on runtime prediction + Waittime prediction

The Local Scenarios

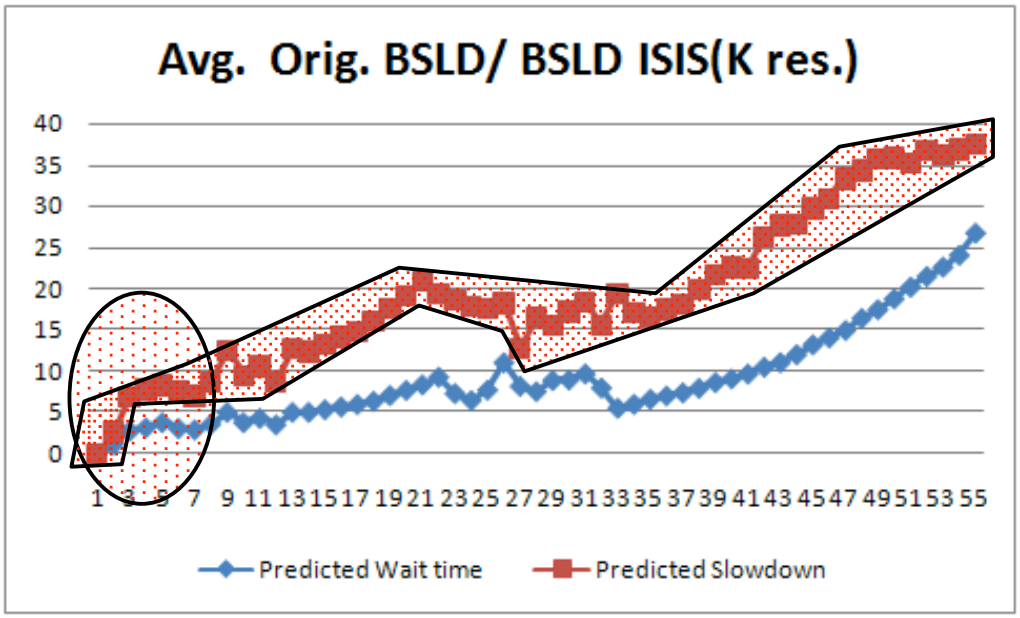
Francesc Guim Bernat

Metric:	Wait time		Slowdown	
Center	**Avg**	95_{th}	**Avg**	95_{th}
CTC-SP2	5249	29586	7,76	39,01
LCG	434,12	4320	4,3	23,32
DAS2-fs0	22,68	135	1,11	1,69
DAS2-fs1	5576	43414	11,71	21,18
DAS2-fs2	29594	99109	6,33	14,44
DAS2-fs3	4,52	100	1,03	3,23
DAS2-fs4	39053	192140	221	934,33
HPC2N	23980	87607	72,05	299,5
KTH-SP2	8864	54222	74,46	571,5
LANL-CM5	126565	308231	1364	4061
LPC	133	1323	1,23	3,42
Atlas	1993	14217	3,18	12,97
Thunder	18891	47758	138	366,8
BLUE	12383	27644	68,80	164,2
Par	453,12	12000	7,32	18,42
OSC	1233,32	25433	5,443	24,43
SDSC-SP2	116,12	1233	1,45	4,22
NASA	232,45	2133	2,43	10,43
Sharnet	649	4432	43,6	749
All	18198	29345	135,5	653

- Wait time
  - Minimum Avg. : 5 secs (DAS2-fs3)
  - Avg. Avg.: 18198 secs
  - Maximum Avg. : 126565 secs (LANL-CM5)
- Slowdown
  - Minimum Avg. : 1,03 (DAS2-fs3)
  - Avg. Avg.: 135
  - Maximum Avg. : 1364 (LANL-CM5)

Slowdown Improvement

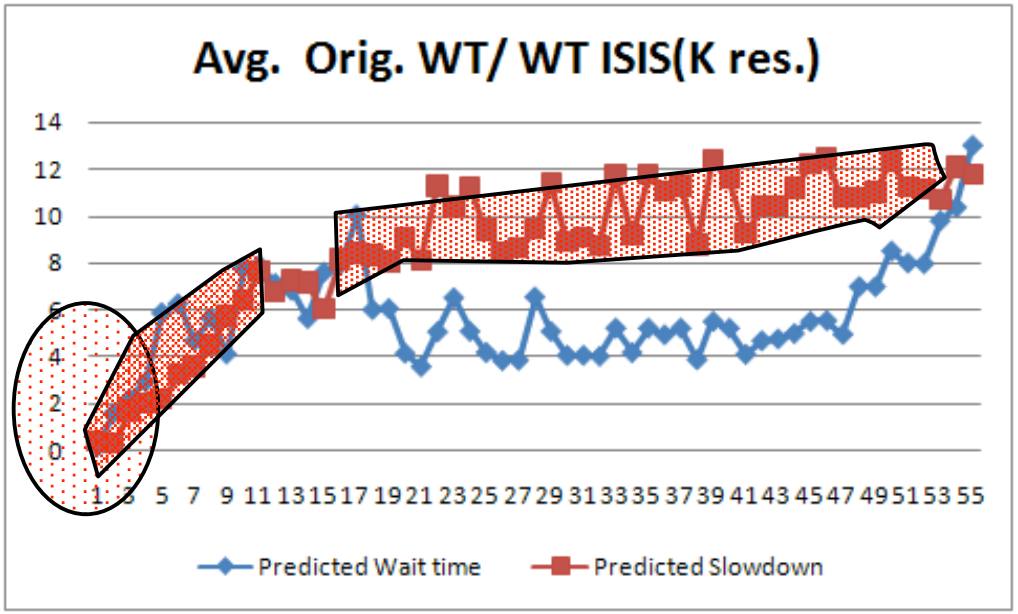
Francesc Guim Bernat



- Improvements from
  - Less-Waittime → k=3
  - Less-Slowdown → k=2
- Qualitative improvements
  - Less-Waittime → k>13
  - Less-Slowdown → k>4
- The Less-Slowdown shows better results.
- A good trade-off between k and slowdown
  - k=5

Wait time

Francesc Guim Bernat



- Improvements from
  - Less-Waittime → k=2
  - Less-Slowdown → k=3
- Qualitative improvements
  - Less-Waittime → k>5
  - Less-Slowdown → k>9
- The Less-Slowdown shows better results from k>10
- The Less-Waittime shows better results from k<10
- A good trade-off between k and slowdown
  - k=6

Outline

- The XtreemOS Project
- The scenario and its challenges
- The ISIS-Dispatcher
- Including the ISIS-Dispatcher in the XOS
- Evaluation
- Conclusions & Future Work

Conclusions and Future Work

- We have presented how the ISIS-Dispatcher can be used in XOS
  - Using prediction system
  - Using the ADS system
  - Providing good Slowdown and Wait time performance
- We have shown the impact of the ADS
  - In general, from K=3 we have good metrics values
  - In general, from k>10 we have a qualitative improvement
- Future work must include
  - Consideration of *on-fly* submissions
  - Consideration of reservations
  - Consideration of non centralized prediction techniques