EASIS: An Optimized Information Service for High Performance Computing Environment

Can Wu, Xiaoning Wang, Haili Xiao, Rongqiang Cao, Yining Zhao, Xuebin Chi
Computer Network Information Center, Chinese Academy of Sciences
Monday May 21. 2018

contents

01

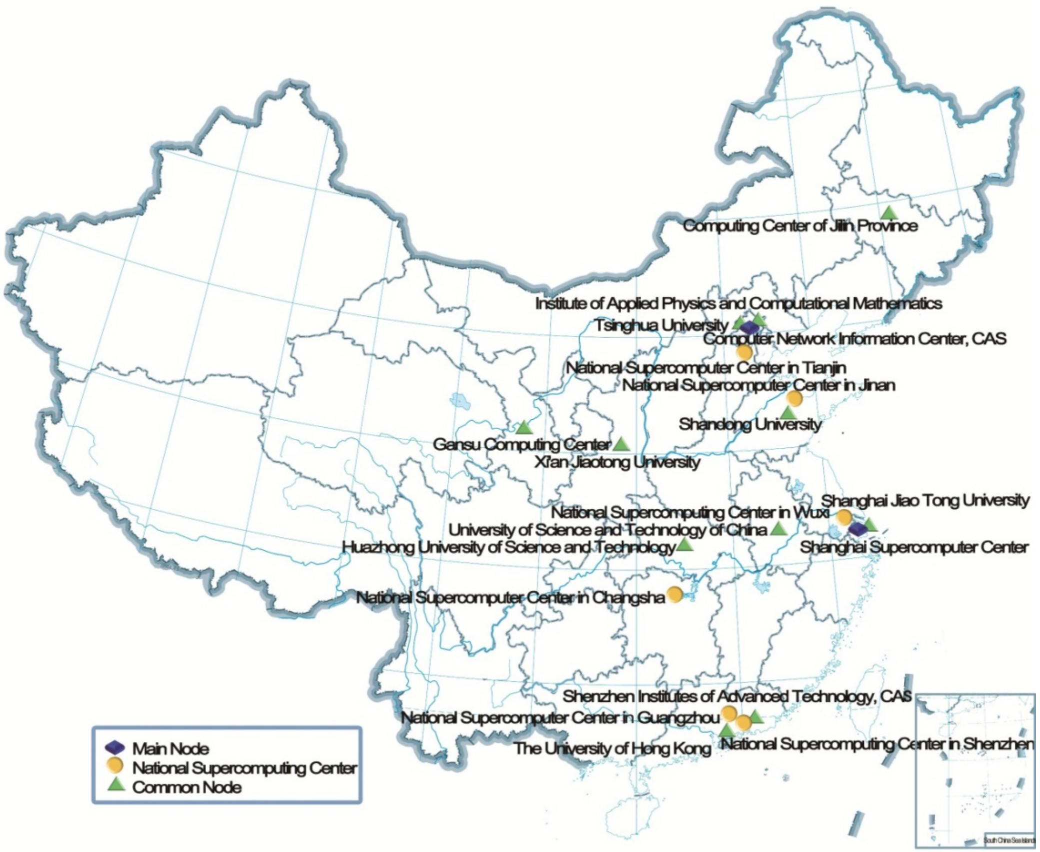
Background

- The High Performance Computing Environment in China（CNGrid）

INTRODUCTION-CNGRID

19 sites, 200PF+167PB

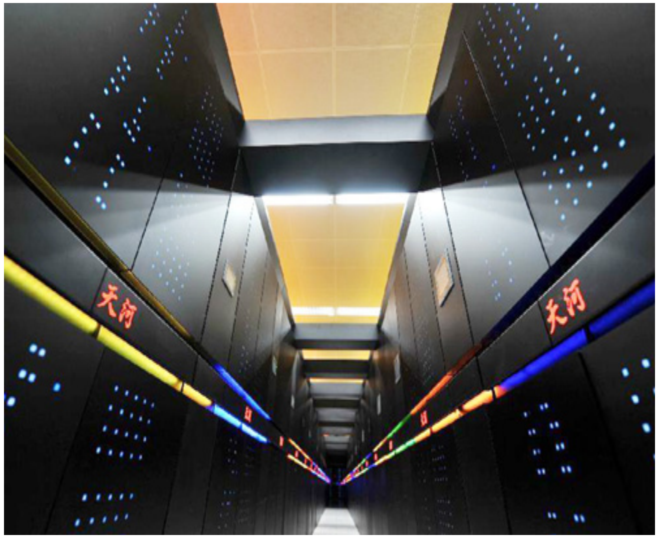
- CNIC (Beijing, major site)
- SSC (Shanghai, major site)
- NSCCWX (Wuxi)
- NSCCGZ (Guangzhou)
- NSCCTJ (Tianjin)
- NSCCSZ (Shenzhen)
- NSCCCS (Changsha)
- NSCCJN (Jinan)
- Tsinghua University (Beijing)
- IAPCM (Beijing)
- USTC (Hefei)
- XJTU (Xi’an)
- SJTU (Shanghai)
- SIAT (Shenzhen)
- HKU (Hong Kong)
- SDU (Jinan)
- HUST (Wuhan)
- GSCC (Lanzhou)
- JLCC(Changchun)



CNGrid HPC

- Sunway TaihuLight
- #1 TOP 500, 2016-now
- Sunway processor: SW26010
- 125 PFlops, 10,649,600 cores
- 15,371 kW
- Wuxi

CNGRID HPC



- Tianhe-2
- #1 TOP 500, 2013-2015
- 55 PFlop/s, 3,120,000 cores
- 17,808.00 kW
- Guangzhou

CNGrid HPC

- Tianhe-1A
- #1 TOP 500, 2010
- 4.7 PFlop/s
- Tianjin

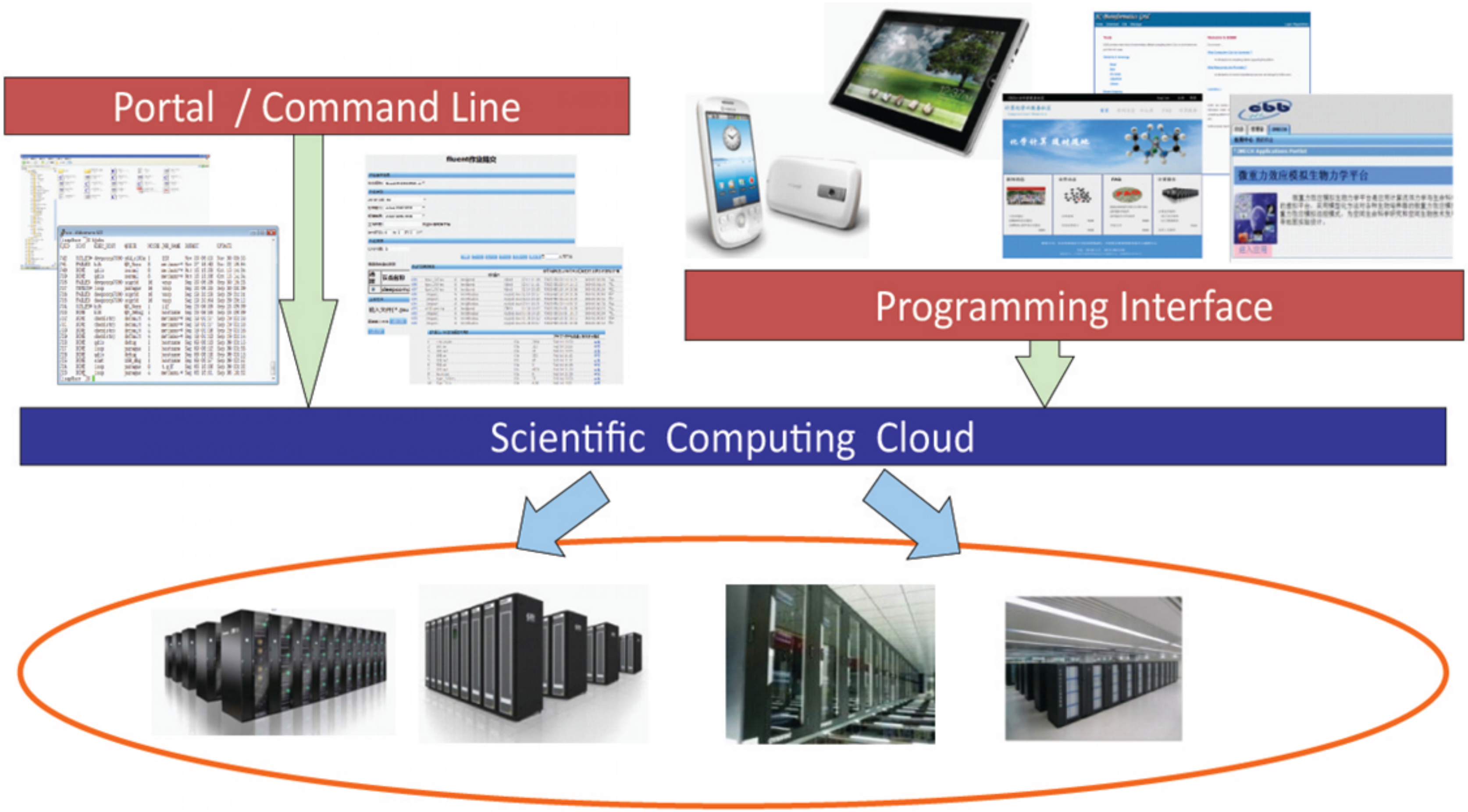
- Sunway Blue Light
- #14 TOP 500, 2011
- ShenWei processor
- 1 TFlop/s
- Jinan

- Nebulae
- #2 TOP 500, 2010
- 2.98 PFlops/s
- Shenzhen

- Era (CNIC)
- 2.3 PFlop/s
- Beijing

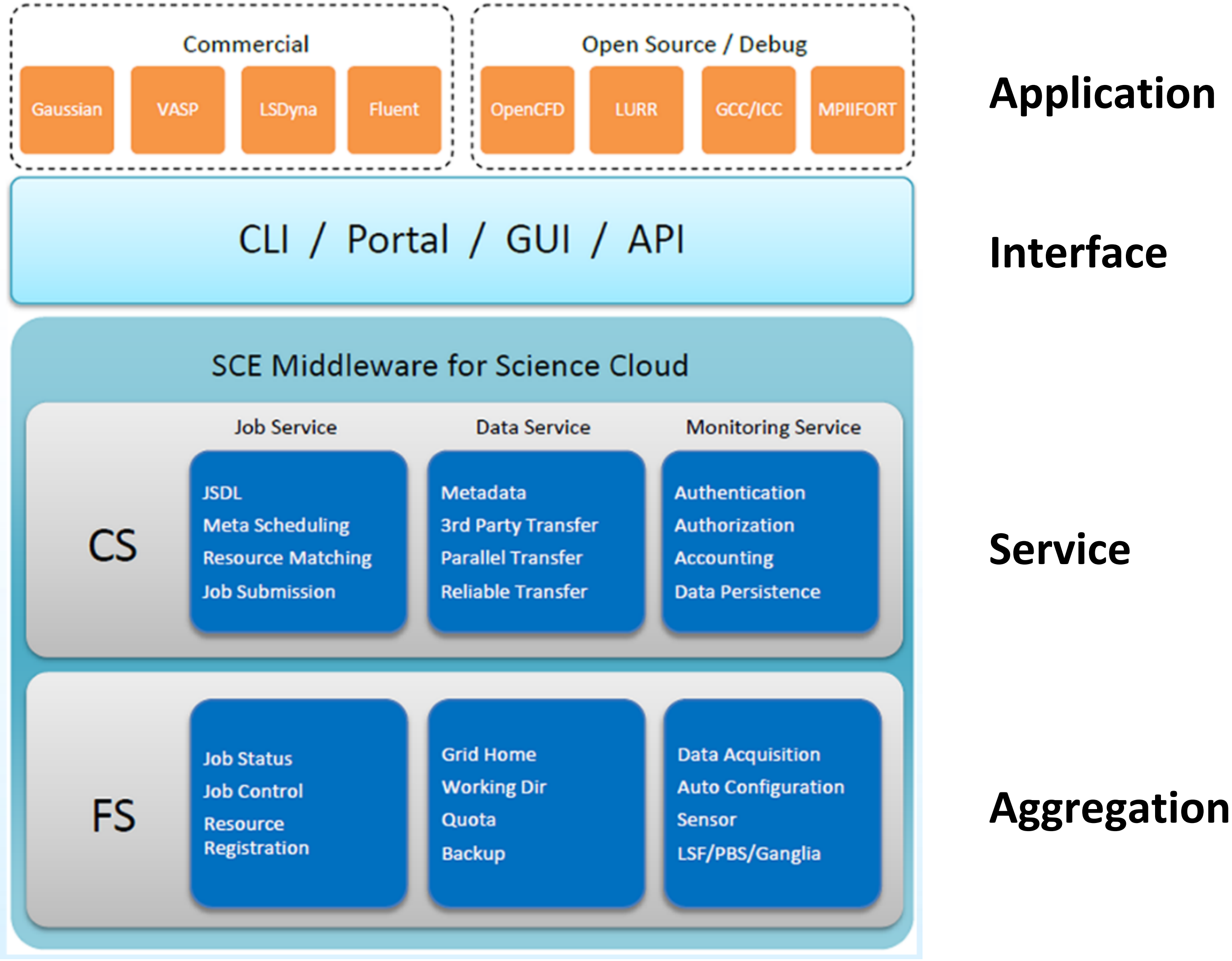
- Dawning 5000A
- #11 TOP 500, 2008
- 233.5 Tflop/s
- Shanghai

INTRODUCTION



Unified Service, Schedule and Management

SCE Middleware



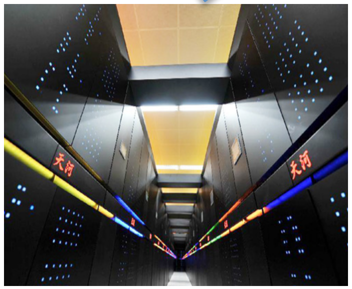
02 Problem and Requirements

- Problem
- Requirements

Problem

233.5 Tflop/s

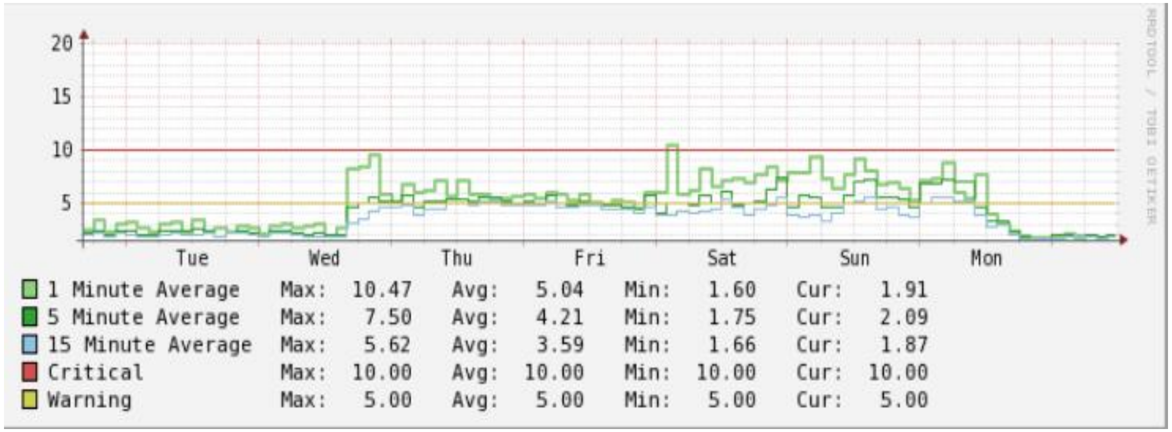
4.7 PFlop/s



55 PFlop/s

125 PFlops

Exascale



REQUIREMENTS

Efficiency: The information service should extract and transfer all the resource information rapidly in the system.

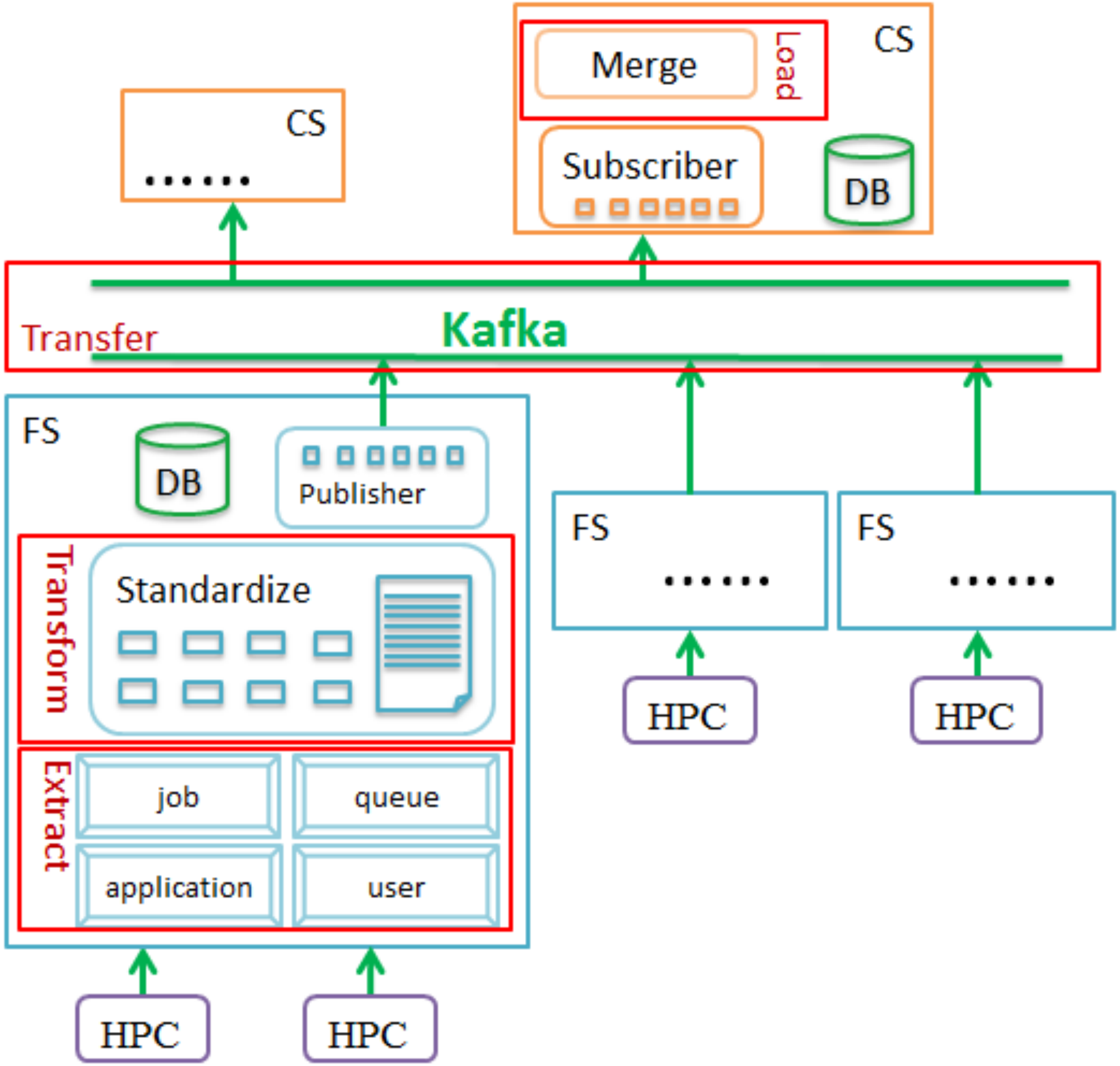
High Availability: The resource information in SCE should not be lost when the network is down off or unstable.

Simplicity: The structure of resource information service should be simple and the interfaces are easy to use.

03 The Structure and Key Technologies

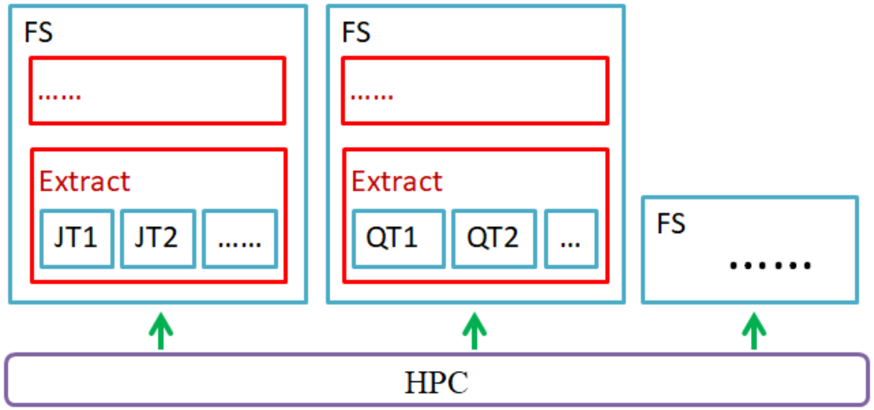
- Structure
- Key Technologies
  - Extracting information concurrently
  - Standardizing information
  - Merging information
  - Fault tolerance

Structure


CS
Merge
Load
Subscriber
DB
CS
....
Transfer
Kafka
FS
DB
Publisher
Transform
Standardize
Extract
job
queue
application
user
FS
....
FS
....
HPC
HPC
HPC
HPC

Key Technologies

— Extracting information concurrently



$$x = R \times \left\lfloor \frac{W \times N}{\alpha C + \beta S + \gamma I^{-1}} \right\rfloor$$

- *x:* the number of concurrent threads.
- *R, α, β, γ:* regulatory factors.
- *W:* the amount of resource information.
- *S:* system load.
- *C:* the number of network connections.
- *N:* the number of cores.
- *I:* the computer hard disk I/O speed.

Key Technologies

— Standardizing information

Information Type	Standard Format of Information
job	{“GID”:“GID”,“status”:“status”,“utime” :“utime” }
queue	{“ID”:“ID”,“hpcname”:“hpcname”,“njobs”:“njobs, “pendjobs”:“pendjobs”,“runjobs”:“runjobs”, “status”:“status”}
application	{“ID”:“ID”,“hpcname”:“hpcname”,“applicationname”:“applicationname”,“version”:“version”, “description”: “description”}
user	{“username”:“username”,“hpcname”:“hpcname”}

Job System	Job Status	Symbol
LSF	PEND, PSUSP, WAIT	PEND
	RUN, USUSP, SSUSP	RUN
	DONE	DONE
	EXIT, UNKWN, ZOMBI	EXIT
	*	ERROR
PBS	Q	PEND
	R	RUN
	C, E	DONE
	H	HOLD
	*	ERROR
SLURM	PD, CF	PEND
	R, CG	RUN
	CD	DONE
	CA, F, NF, PR, TO	EXIT
	S	HOLD
	*	ERROR

Key Technologies

— Merging information

Invalid-Update

{"GID": "9632587419632587411", "status" : "PEND", "utime" : "1512543239"},
{"GID": "9632587419632587411", "status" : "RUN", "utime" : "1512543267"},
{"GID": "9632587419632587411", "status" : "DONE", "utime" : "1512543299"}

{"GID": "9632587419632587411", "status" : "DONE", "utime" : "1512543299"}

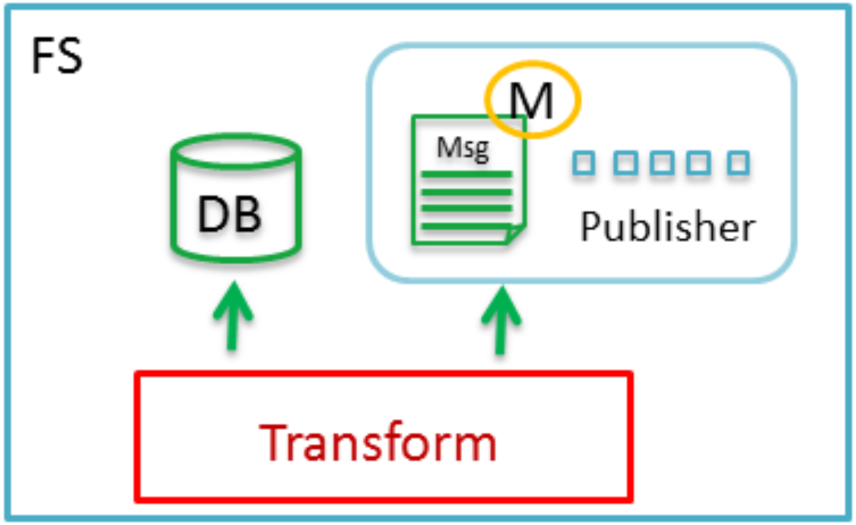
Merged-Update

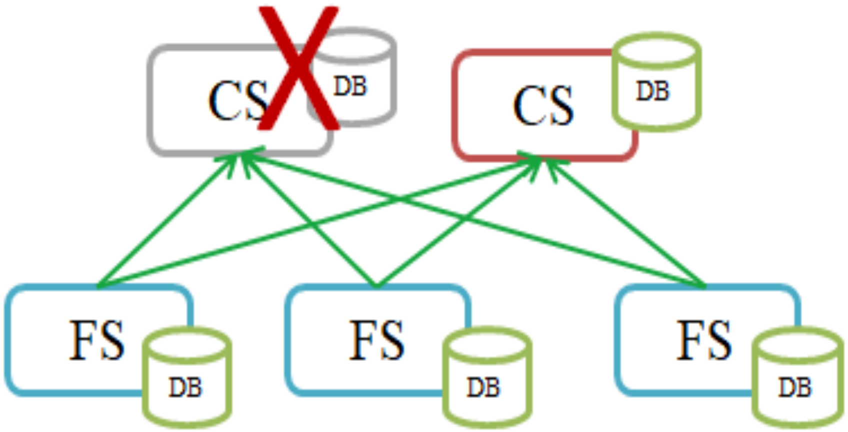
{"GID": "9632587419632587411", "status" : "PEND", "utime" : "1512543239"},
{"GID": "9632587419632587412", "status" : "PEND", "utime" : "1512543239"}

{"GID":"9632587419632587411","9632587419632587412","status":"PEND","utime":"1512543239"
}

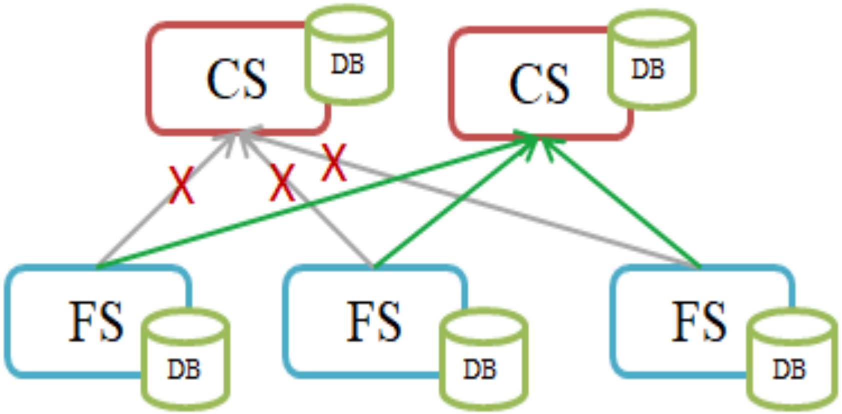
Key Technologies

- Fault tolerance





Server-Fault-Tolerant



Network-Fault-Tolerant

04 Evaluation

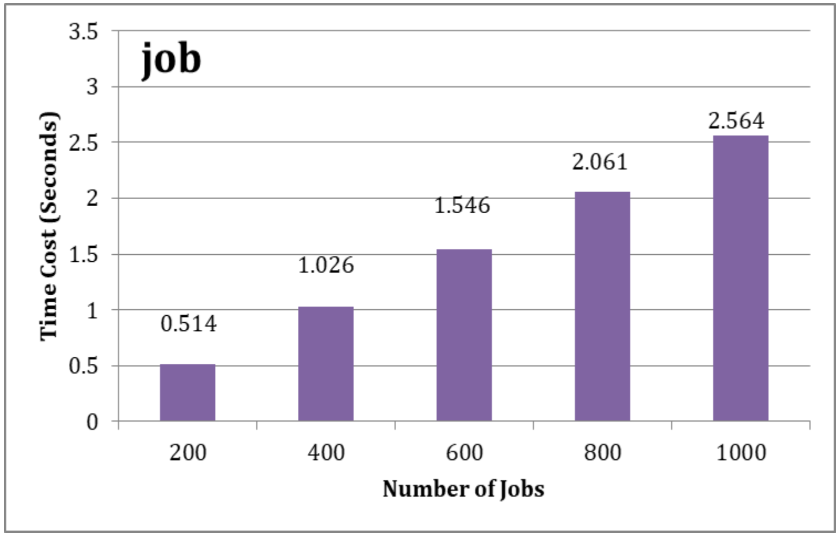
- Efficiency
- Availability

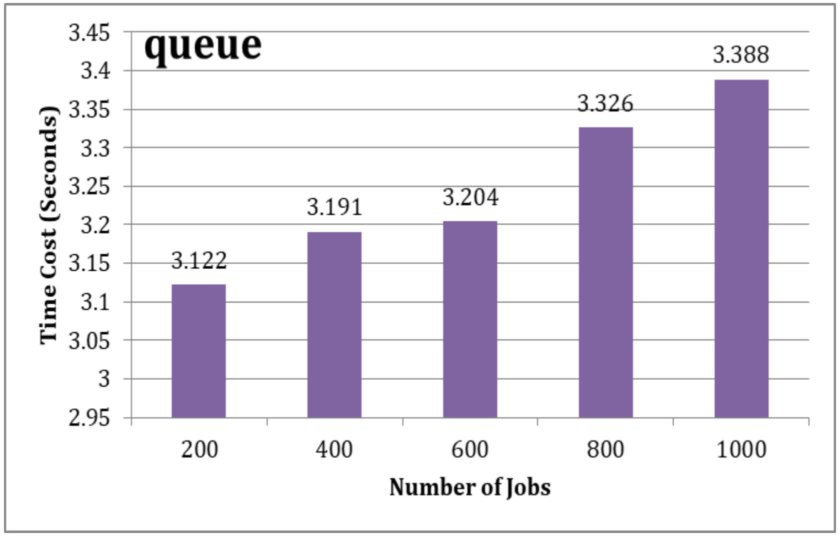
Efficiency Experiment Environment

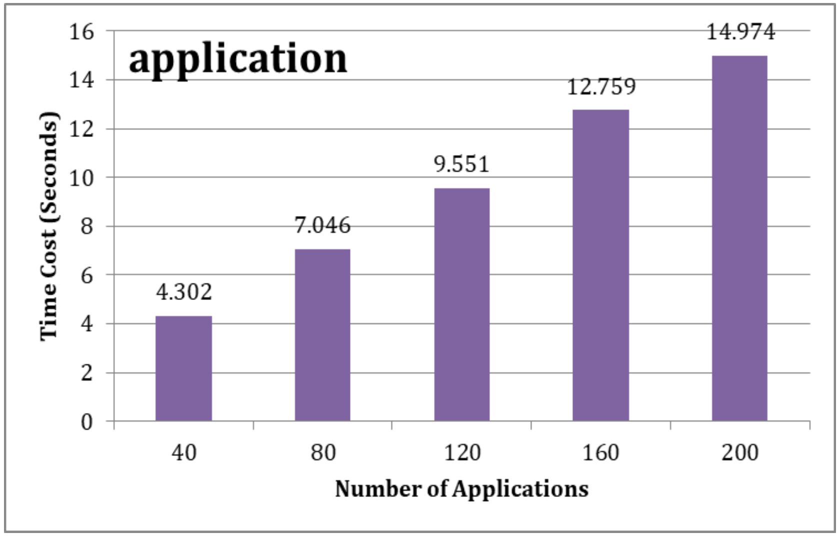
Item	Configuration
CPU	Intel® Xeon® CPU E5-2680 v3 @ 2.50GHz, 1CPU
Cache	30720 KB
Memory	2G
Hard disk	40G
OS	CentOS release 6.9, 64 bit
DB	MySQL 5.1.73

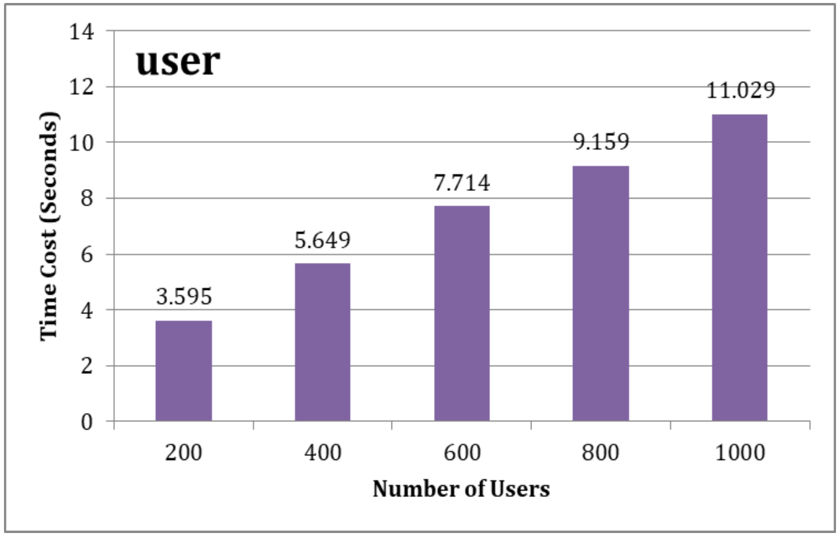
EVALUATION

Efficiency



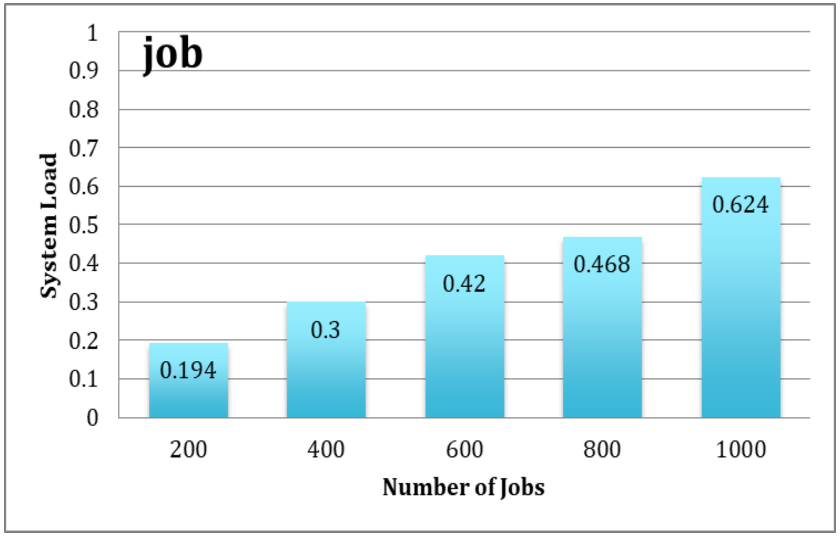


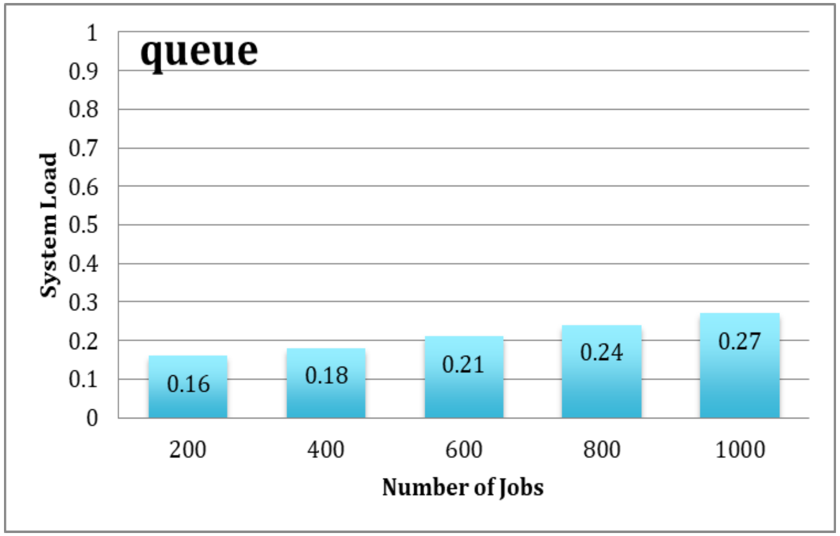


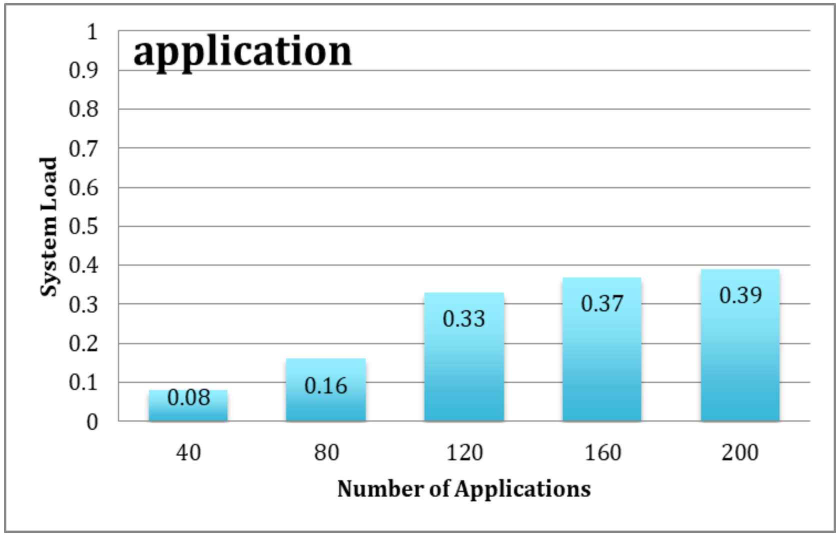


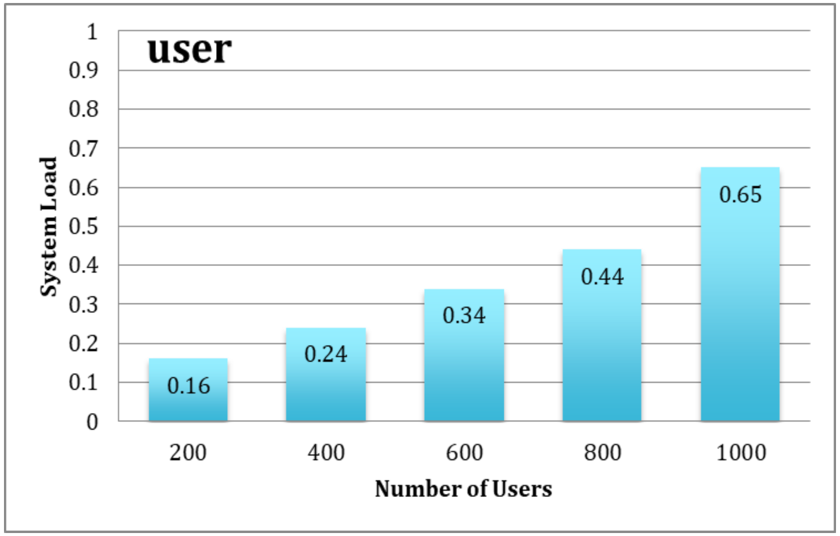
Evaluation

Efficiency



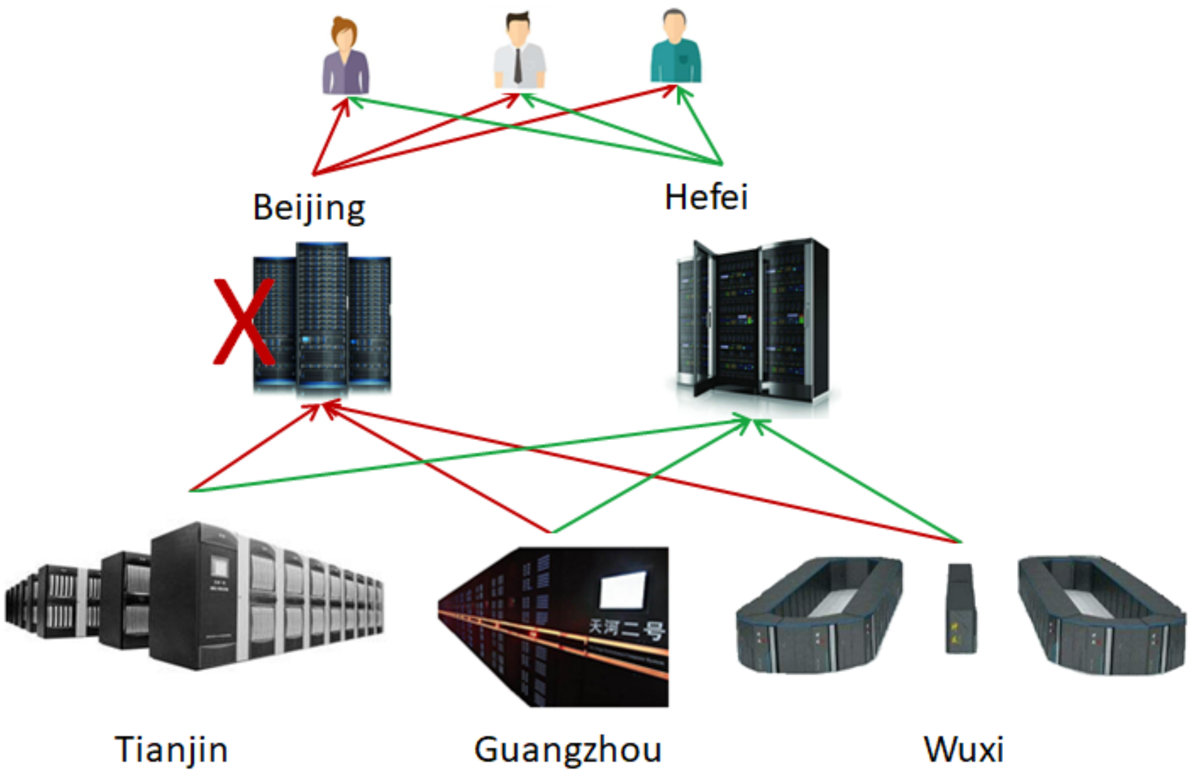






Evaluation

Availability



05

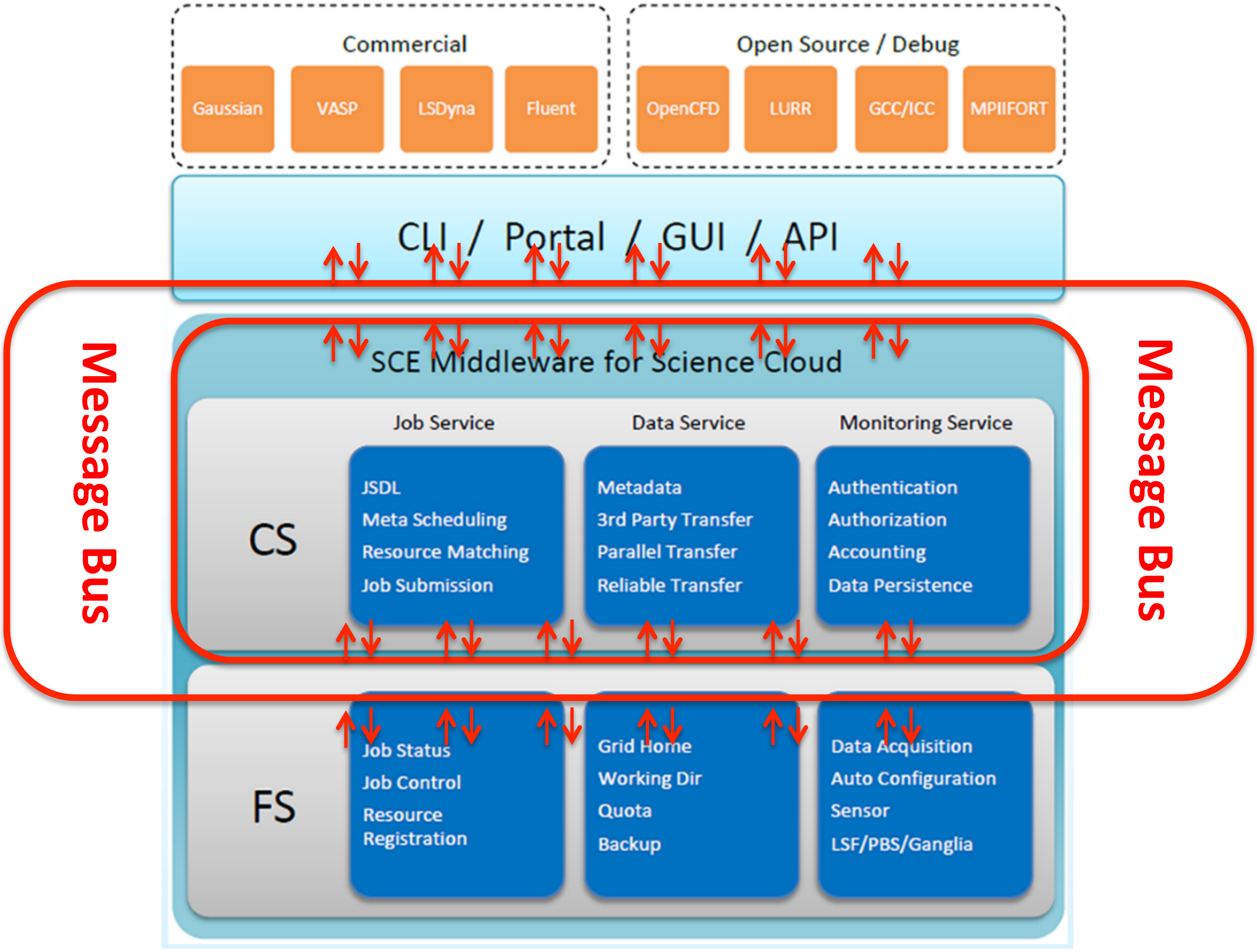
Conclusions and Future Work

- Conclusions
- Future Work

CONCLUSIONS

- EASIS can process 1000 pieces of key resource information with less than 4s and low system load.
- The fault-tolerant system makes the EASIS more available and easier to use.

Future Work


Commercial
Gaussian
VASP
LSDyna
Fluent
Open Source / Debug
OpenCFD
LURR
GCC/ICC
MPIIFORT
CLI / Portal / GUI / API
SCE Middleware for Science Cloud
Message Bus
Message Bus
Job Service
Data Service
Monitoring Service
CS
JSDL
Meta Scheduling
Resource Matching
Job Submission
Metadata
3rd Party Transfer
Parallel Transfer
Reliable Transfer
Authentication
Authorization
Accounting
Data Persistence
FS
Job Status
Job Control
Resource Registration
Grid Home
Working Dir
Quota
Backup
Data Acquisition
Auto Configuration
Sensor
LSF/PBS/Ganglia

06 Acknowledgement

- Our Team
  - Xuebin Chi, Haili Xiao, Rongqiang Cao
  - Xiaoning Wang, Yining Zhao, Shasha Lu, Rong He
  - Can Wu, Xiaodong Wang, Yimeng Han
- HPBDC

Computer Network Information Center
Chinese Academy of Sciences

Thank you!

http://www.cngrid.org

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences