NEXTGenIO: Resource Requirement Specification for Novel Data-aware and Workflow-enabled HPC Job Schedulers

Manos Farsarakis
@efarsarakis
e.farsarakis@epcc.ed.ac.uk
EPCC, The University of Edinburgh
Hi, I’m Manos!
NEXTGenIO summary

Project

• Design, develop, and exploit HPC and HPDA system with NVRAM in compute nodes
• 36 month duration
• €8.1 million
• Approx. 50% committed to hardware development
• http://www.nextgenio.eu/
• This project has received funding from the European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement no. 671951.

Partners
Our objectives

• Hardware platform prototype
  • Demonstrating the prototype’s broad applicability for both HPC and data centric applications

• Exascale I/O investigation
  • Understanding how best to exploit NVRAM

• Systemware development:
  • Producing the necessary software to enable Exascale application execution on the hardware platform

• Application co-design
  • Understanding individual application I/O profiles and typical I/O workloads on shared systems running multiple different applications
Old System

Network

Filesystem
New System (1)
I/O Performance

- https://www.archer.ac.uk/documentation/white-papers/parallelIO-benchmarking/ARCHER-Parallel-IO-1.0.pdf
I/O

Individual I/O Operation

I/O Runtime Contribution
Age old question...
Questions for you

• How do YOU do I/O?
• How much I/O do YOU do?

But more importantly…

• How do you WANT to do I/O?
• How much I/O would you WANT to do?
Types of things we are thinking about...

- Often read, never write files
- Frequently used files
- Temporary runtime files
- Disaster recovery files
- Workflows (which often include the above topics with renewed importance)
Workflows

Resources

Time

Job 1

Job 2

Job 3
Workflows

Resources

Time

Job 1

Job 2

Job 3

Read-in, write-out
Temporary files
Workflows: Data Aware (1)

Resources

Job 1

Job 2

Job 3

Time

Resources

Read-in, write-out
Temporary files
Workflows: Data Aware (2)

Resources

Time

Job 1

Job 2

Job 3

Read-in, write-out
Temporary files
The Problem:

Data Aware Scheduler needs information about the data!
### JOB RESOURCE REQUIREMENT EXAMPLE FILE###
### STANDARD JOB VOCABULARY ###

#JOB_NAME: JOB2    ### ( OF 3 JOB WORKFLOW)
#QUEUE: STANDARD
#WALLTIME: 00:30:00
#NUM_NODES: 2
#PROCS_PER_NODE: 3
#DEPENDENT_ON: JOB
### AUGMENTED JOB RESOURCE REQUIREMENTS SPEC ###
### NEW VOCABULARY ###

#HPS: [
[../JOB1/ON_NODE_FILES/checkpoint%[0-9]+ %.out, 10MB, IN],
[../JOB2/READ_ONLY/initial_conditions% [0-9]+ %.dat, 2MB,IN],
[../JOB2/SHARED/param_dictionary.dat, 50MB, READ_ONLY],
[../JOB2/LOCAL/savedParams%[0-9]+ %.out, 5MB, LOCAL],
[../JOB2/OUTPUT/checkpoint%[0-9]+%.out, 19MB,OUT]
]

#PM_FILE: "/PATH/TO/JOB2/mappings.dat"

aprun -n 6 -N 3 ./myApp --input=/PATH/TO/JOB1/ON_NODE_FILES
Summary

• NEXTGenIO developing a **full** hardware and software solution

• Data-Aware-Scheduler development has shown us that current job descriptions are not enough

• We have introduced JRRS as a means to bridge this gap

• Development is in initial stages: We welcome input!