Enabling Large-scale Scientific Workflows on Petascale Resource Using MPI Master/Worker

Mats Rynge, Gideon Juve, Karan Vahi, Gaurang Mehta, Ewa Deelman
Information Sciences Institute, University of Southern California

Scott Callaghan, Philip J. Maechling
Southern California Earthquake Center, University of Southern California

USC Viterbi School of Engineering Information Sciences Institute

pegasus

SCEC an NSF+USGS center

Funded by the NSF OCI program grants OCI-0722019 and OCI-0943725

Outline

Introduction

Pegasus Workflow Management System

pegasus-mpi-cluster

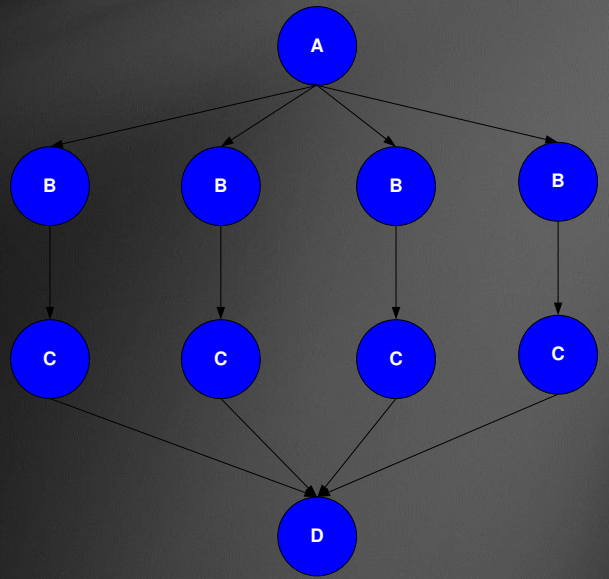
Integration

SCEC Cybershake – Example application

Conclusions

Introduction

- Loosely coupled applications structured as scientific workflow, containing mix of parallel and serial tasks
- Petascale systems are optimized for parallel jobs
- Common solution: Condor glideins for the serial tasks



Glideins on NICS Kraken?

- Cray XT System Environment / ALPS / aprun
  - Login node
  - aprun node
  - Compute node

Approach

- Partition workflow into subgraphs
- Execute partition as a self-contained MPI job

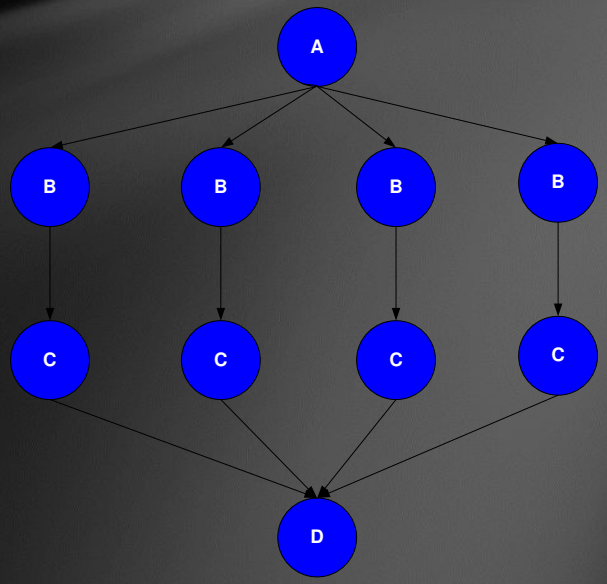
Pegasus Workflow Management System

Abstract Workflows - Pegasus input workflow description

- Workflow "high-level language"
- Only identifies the computation, devoid of resource descriptions, devoid of data locations

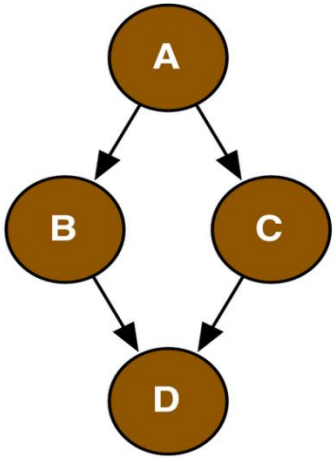
Pegasus

- Workflow "compiler" (plan/map)
- Target is DAGMan DAGs and Condor submit files
- Transforms the workflow for performance and reliability
- Automatically locates physical locations for both workflow components and data
- Provides runtime provenance



pegasus-mpi-cluster

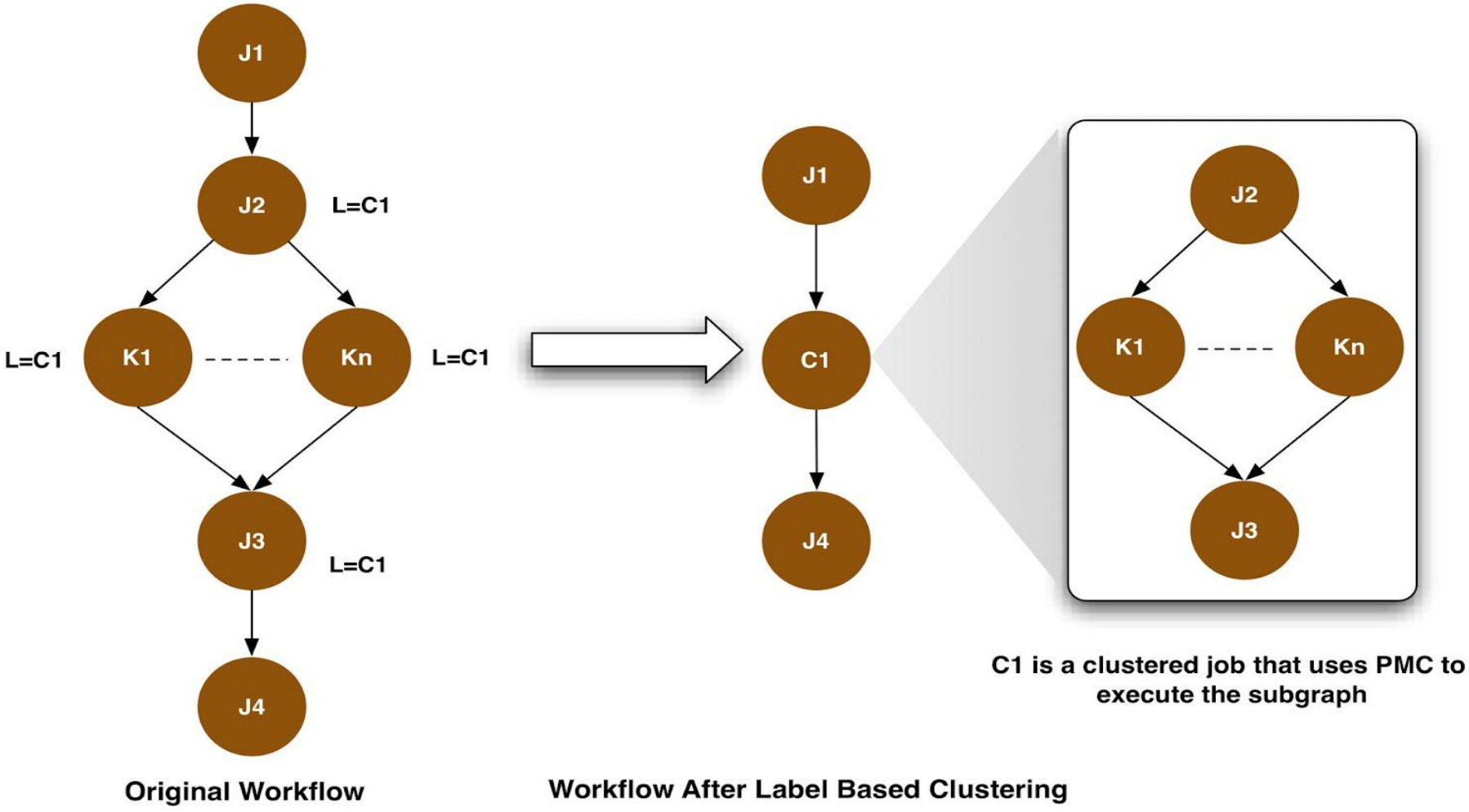
- Master/worker paradigm
- Master manages the subgraph tasks, handing out work to the workers
- Efficient scheduling / handling of input/outputs
- Subgraph described in a DAG-similar format
- Failure management / rescue DAG

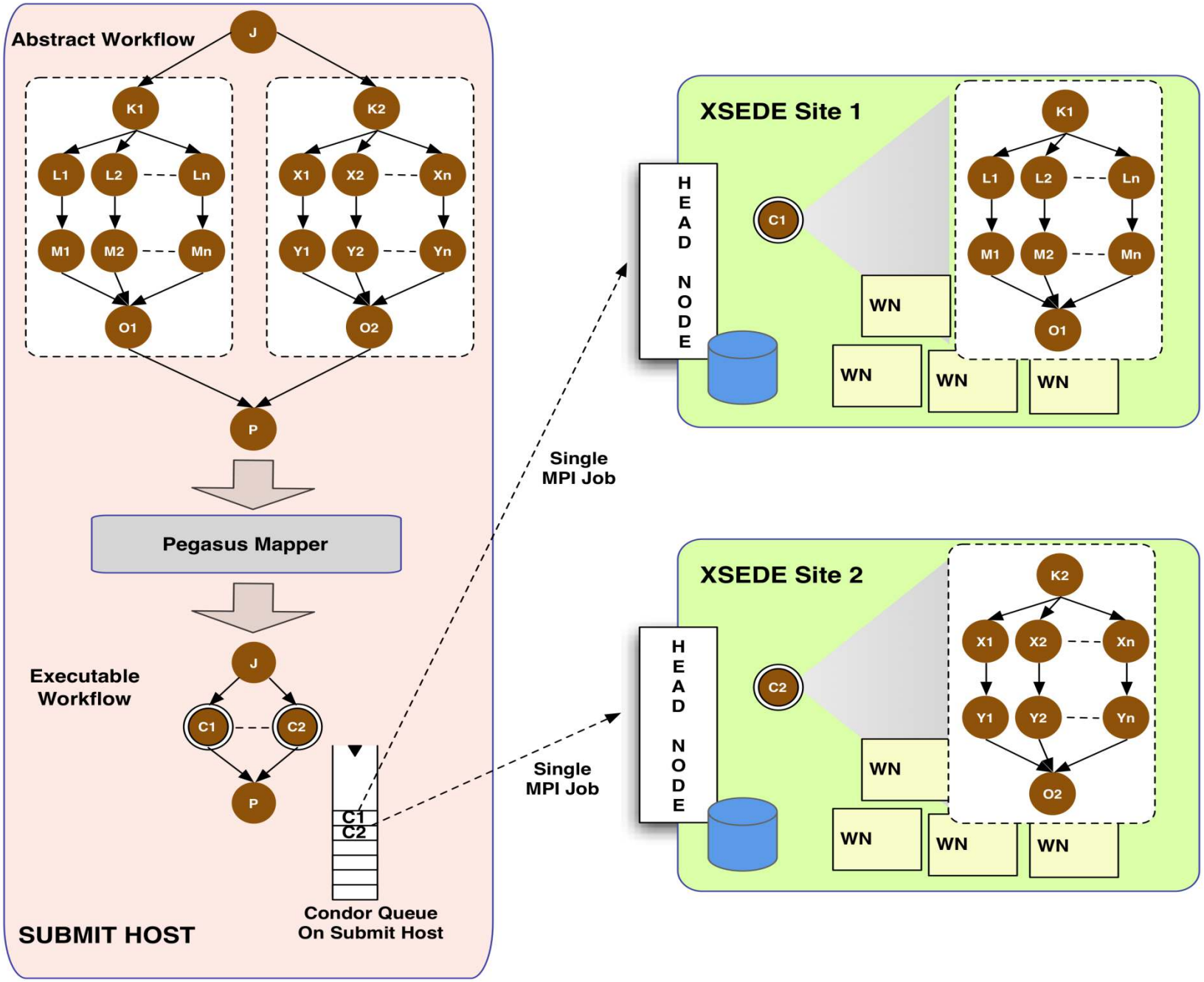


diamond.dag

```
TASK A /bin/echo "I am task A"
TASK B /bin/echo "I am task B"
TASK C /bin/echo "I am task C"
TASK D /bin/echo "I am task D"
EDGE A B
EDGE A C
EDGE B D
EDGE C D
```

Pegasus Mapper / pegasus-mpi-cluster integration




Abstract Workflow
J
K1
K2
L1
L2
Ln
X1
X2
Xn
M1
M2
Mn
Y1
Y2
Yn
O1
O2
P
Pegasus Mapper
Executable Workflow
J
C1
C2
P
C1
C2
Condor Queue On Submit Host
SUBMIT HOST
Single MPI Job
Single MPI Job
XSEDE Site 1
HEAD NODE
C1
WN
WN
WN
WN
K1
L1
L2
Ln
M1
M2
Mn
O1
XSEDE Site 2
HEAD NODE
C2
WN
WN
WN
WN
K2
X1
X2
Xn
Y1
Y2
Yn
O2

an NSF + USGS center

Southern California Earthquake Center

CyberShake

Example Application

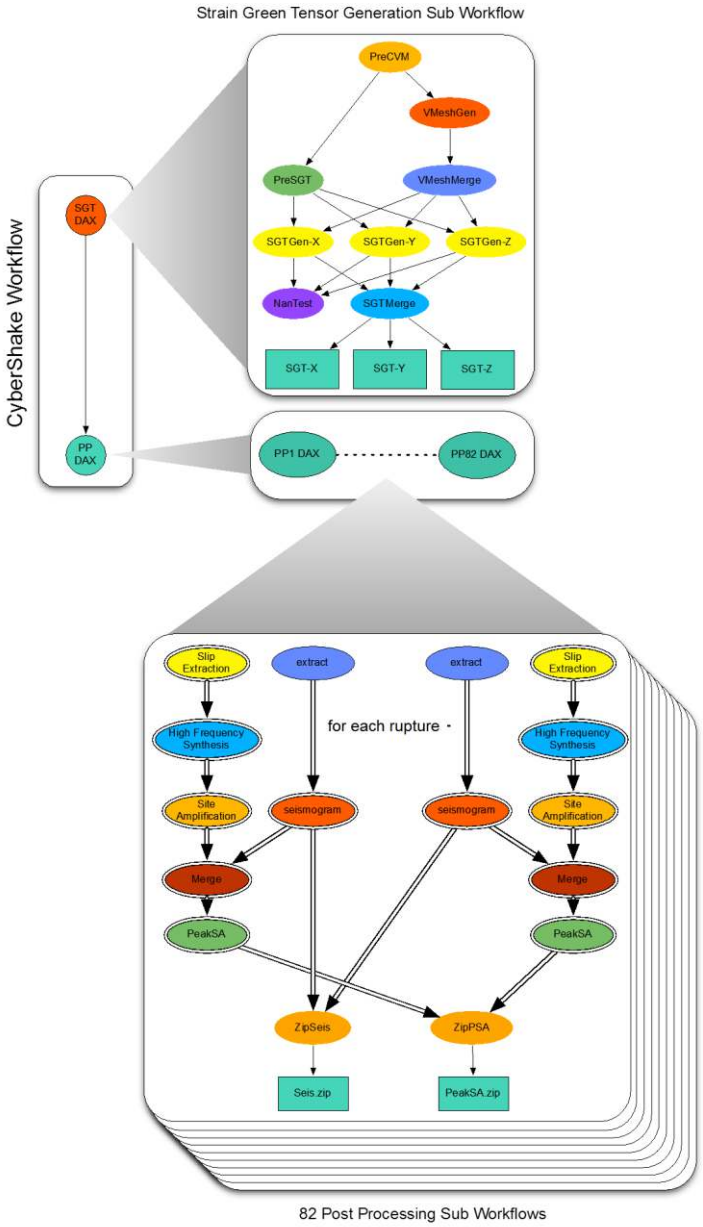
CyberShake

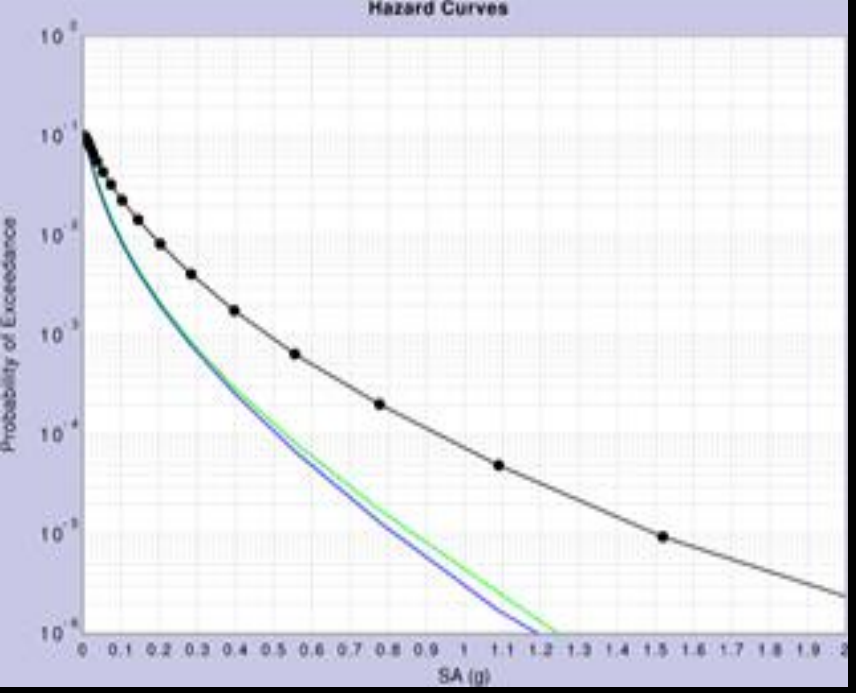
Probabilistic seismic hazard analysis workflow

- How hard will the ground shake in the future?
- Considers a set of possible large earthquakes
- 415,000 earthquakes is typical

Uses Pegasus and Condor DAGMan for workflow management

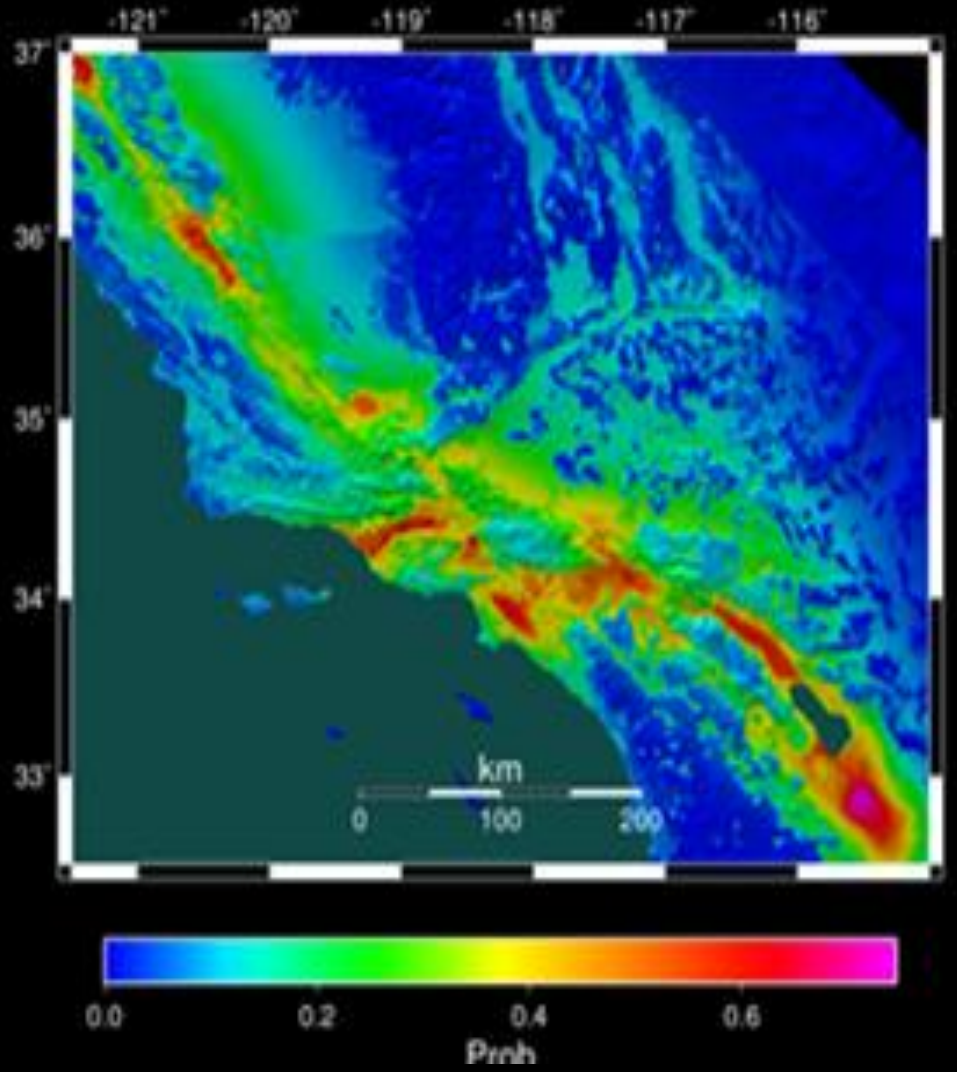
- Hierarchal workflows
- Small set of large parallel jobs
- 840,000 serial jobs, in 78 sub workflows

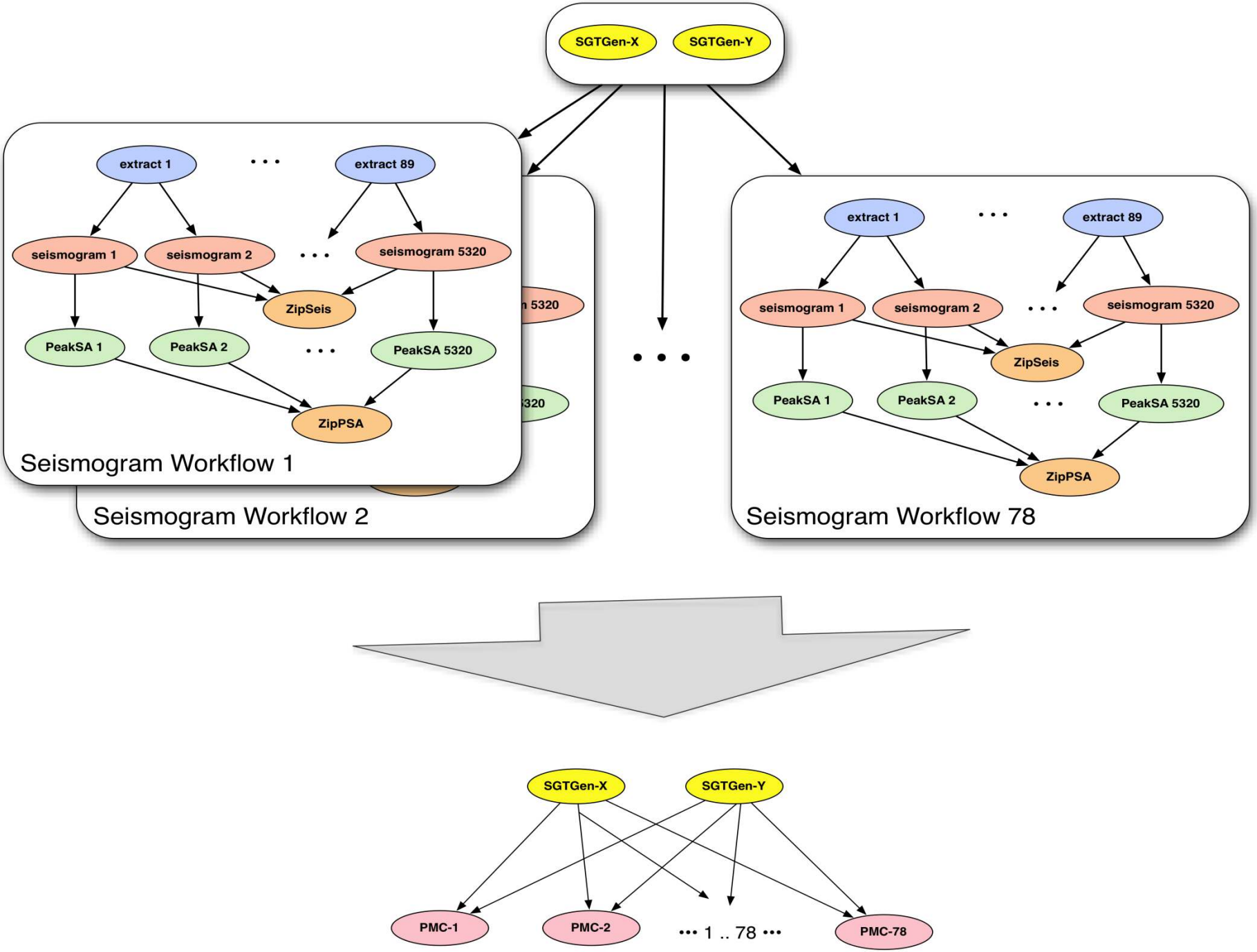




Probabilistic Seismic Hazard Analysis (PSHA) curve. Estimates the probability that earthquake ground motions will exceed some intensity measure.

Set of PSHA curves interpolated creates hazard map for an area




SGTGen-X
SGTGen-Y
extract 1
extract 89
seismogram 1
seismogram 2
seismogram 5320
ZipSeis
PeakSA 1
PeakSA 2
PeakSA 5320
ZipPSA
Seismogram Workflow 1
Seismogram Workflow 2
Seismogram Workflow 78
SGTGen-X
SGTGen-Y
PMC-1
PMC-2
... 1 .. 78 ...
PMC-78

Conclusions and Future Work

- Demonstrated efficient execution of fine-grained workflows on petascale resources by partitioning workflow into MPI master/worker jobs
- Size of partition?
- Size of MPI job?
- Handing tasks with mixed requirements?
  - pegasus-mpi-cluster now considers memory to be a consumable resource

Thank you!

Pegasus: http://pegasus.isi.edu

SCEC: http://www.scec.org