Workflow for extreme-scale systems

Presenter: Michael Wilde wilde@mcs.anl.gov

Mathematics and Computer Science Division
Argonne National Laboratory
University of Chicago/Argonne Computation Institute

Outline

- Overview – context of workflow for science and engineering
- Workflow environments
- Expressing workflows – tools and programming models
- Workflow issues for extreme scale
- IO performance envelopes for workflow
- Expressing workflows in Swift
- Hands-on workflow examples and exercises using Swift
  - Language basics using Swift/K
  - Running Swift/T on Blue Gene/Q
  - Running Swift/K on Tukey

Definitions

- Workflow: the execution of a set of application programs
  - Often for a diverse set of application programs
  - Often with logical and physical dependencies
    - Logical: data dependencies
    - Physical: resource dependencies (space, processor, solution priorities)
  - Scripting is one way to implement workflows (Ad-hoc, Parallel libraries, Swift)
  - Generation of engine-specific input is another (DAGMan, Pegasus, Galaxy, Kepler)
- Scripting: higher-level dynamic programming
  - J. Ousterhout: "Scripting: Higher level programming for the 21st century"
- High throughput computing (HTC)
- Many-task computing (MTC)
- Dataflow
- Data parallel vs. task parallel
  - Workflow is almost always task-parallel at its outer levels
  - SPMD: typified by MPI
  - MPMD: multiple programs, multiple data – more typical of workflow

Many-task Applications

- Many-task Computing applications assemble existing parallel or sequential programs
- Those programs read and write data to a filesystem (but this limitation is being overcome…)
- Applications often have multiple stages
- Task dependencies between stages are in the form of file production and consumption
- Can have very high rates (eg hundreds per second) of very short tasks (minutes down to sub-second)

Slide courtesy of Zhao Zhang

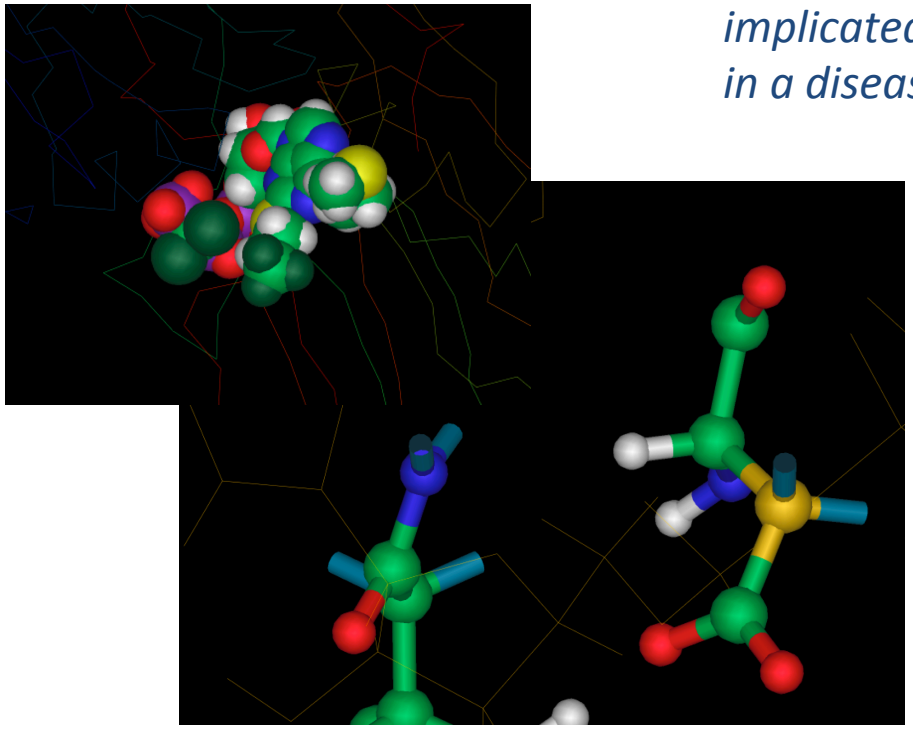
When do you need workflow?

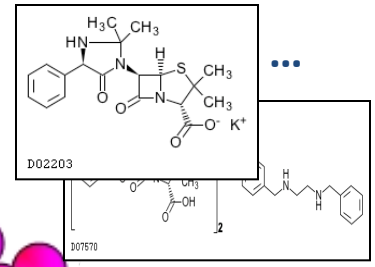
Typical application: protein-ligand docking for drug screening

O(10) proteins implicated in a disease

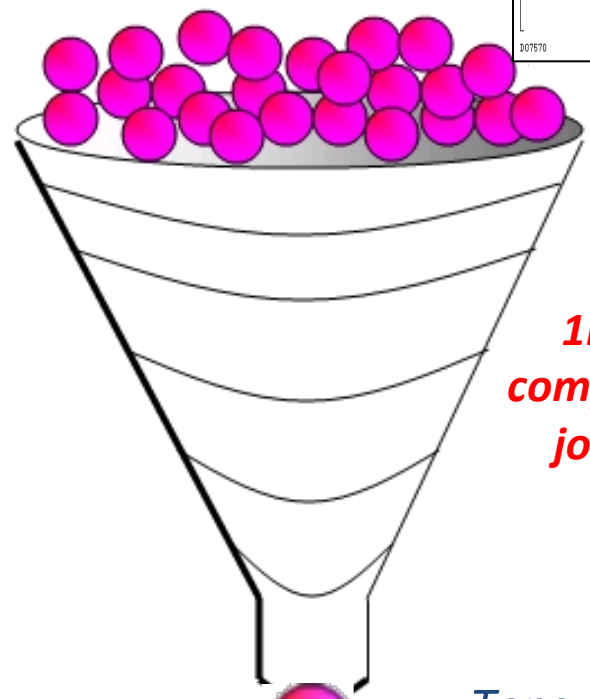
X

O(100K) drug candidates



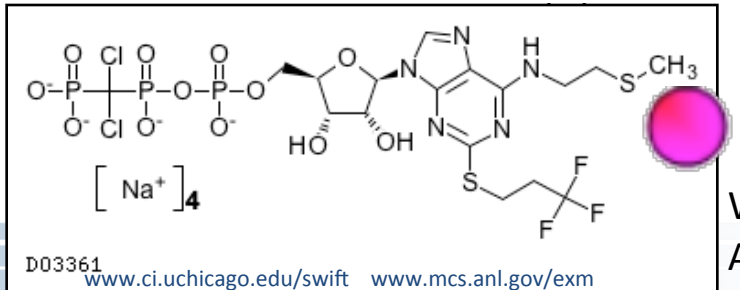


...



1M compute jobs

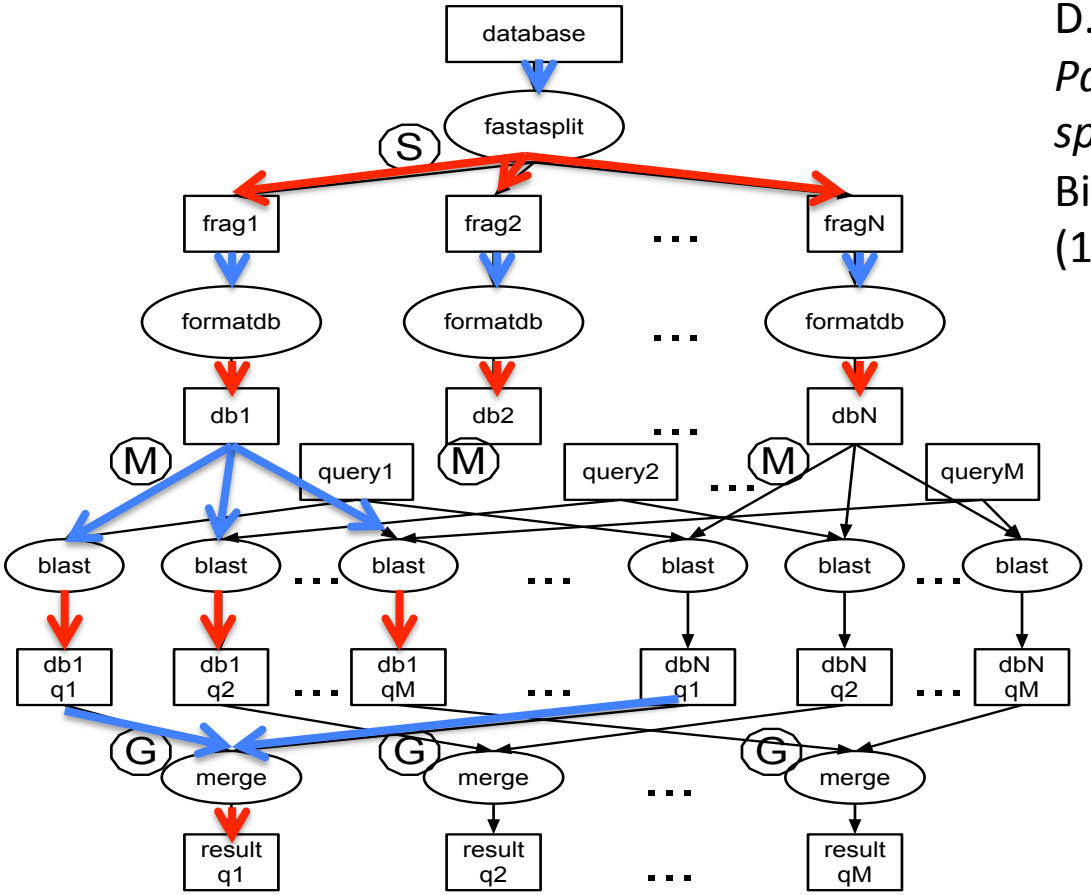
Tens of fruitful candidates for wetlab & APS



Work of M. Kubal, T.A.Binkowski, And B. Roux

Parallel BLAST as a workflow

Original script by D. R. Mathog, *Parallel BLAST on split databases.* Bioinformatics, 19 (14), 2003.

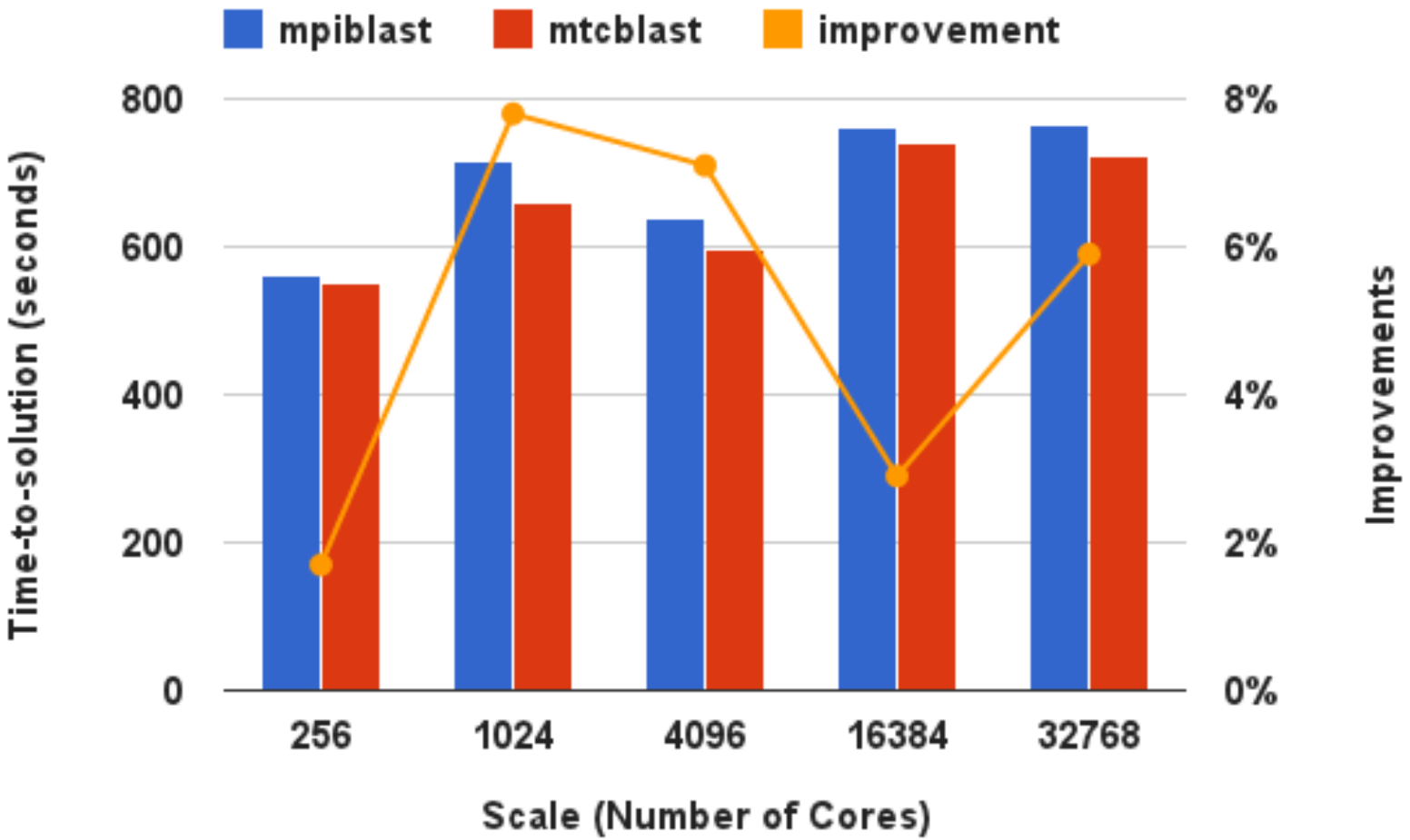


Based on script of D. Matthog by Z.Zhang, L. Gahelha

BLAST STAGE TASKS, INPUTS, OUTPUTS, AND INPUT AND OUTPUT SIZE

Stage	# Tasks	# In	# Out	In (MB)	Out (MB)
fastasplitn	1	1	N	4039	4039
formatdb	N	N	3N	4039	4400
blastp	N*M	N+M	N*M	73*N*M	2.4*N*M
merge	M	N*M	M	2.4*N*M	4.8*M

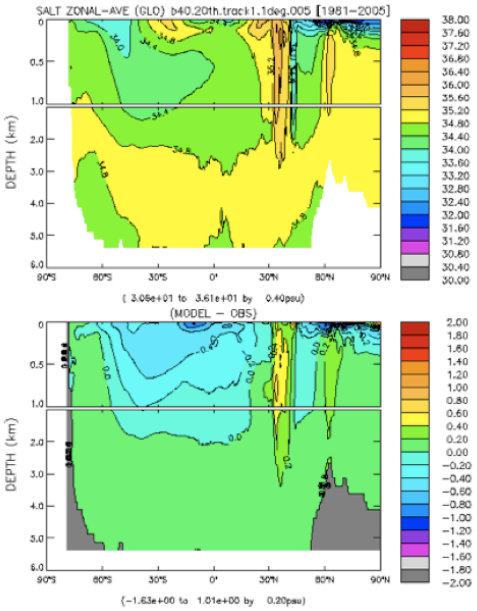
Can workflow scale?
BLAST workflow lags MPI BLAST by ~ 5%

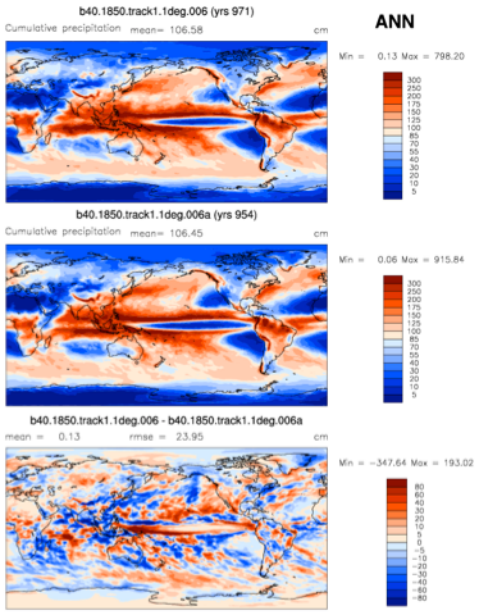


Z. Zhang, D. S. Katz, J. Wozniak, A. Espinosa, I. Foster. "Design and Analysis of Data Management in Scalable Parallel Scripting", Supercomputing 2012.

Analysis & visualization of high-resolution climate models

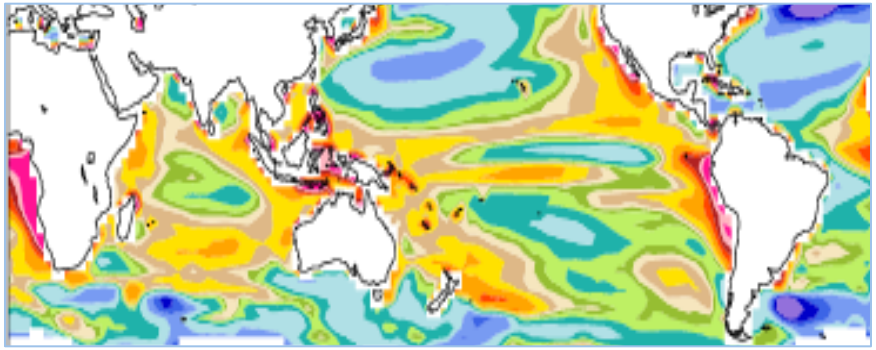
parVis



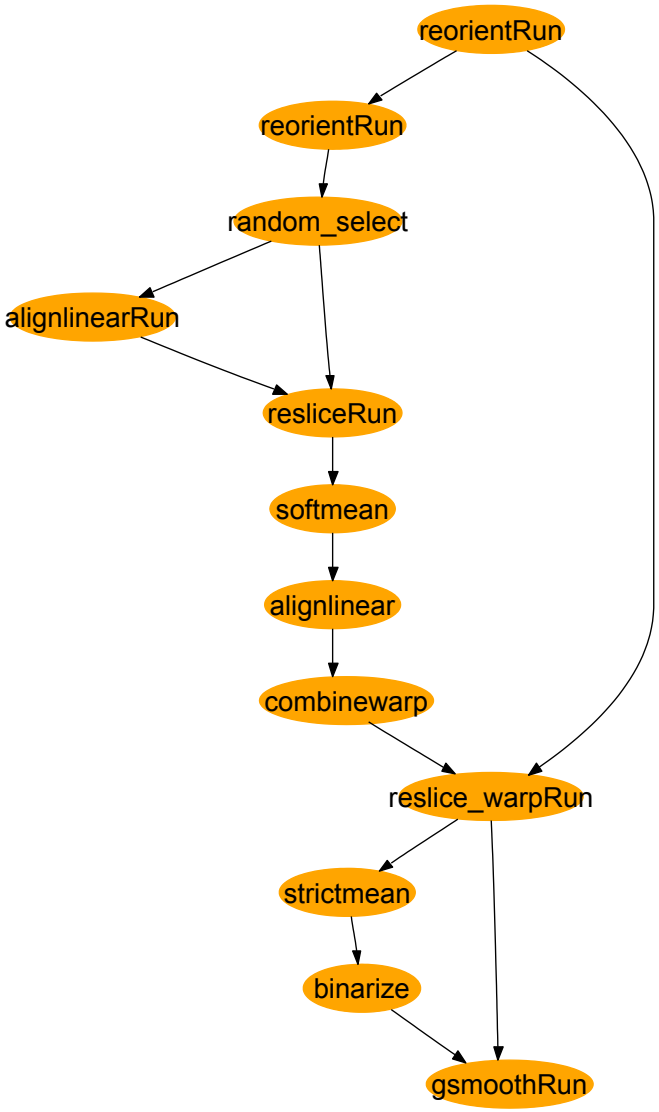


Climate models are continuing to increase both their resolution and the number of variables resulting in multi-terabyte model outputs. This large volume of data overwhelms the series of processing steps used to derive climate averages and produce visualizations. Since many of the tasks in the post-processing sequence are independent, we have applied task-parallel scripting to speed up the post-processing. We have re-written portions of the complex shell script that process output from the Community Atmosphere Model in Swift, a high-level implicitly-parallel scripting language that uses data dependencies to automatically parallelize a workflow. This has resulted in valuable speedups in model analysis for this heavily-used procedure.

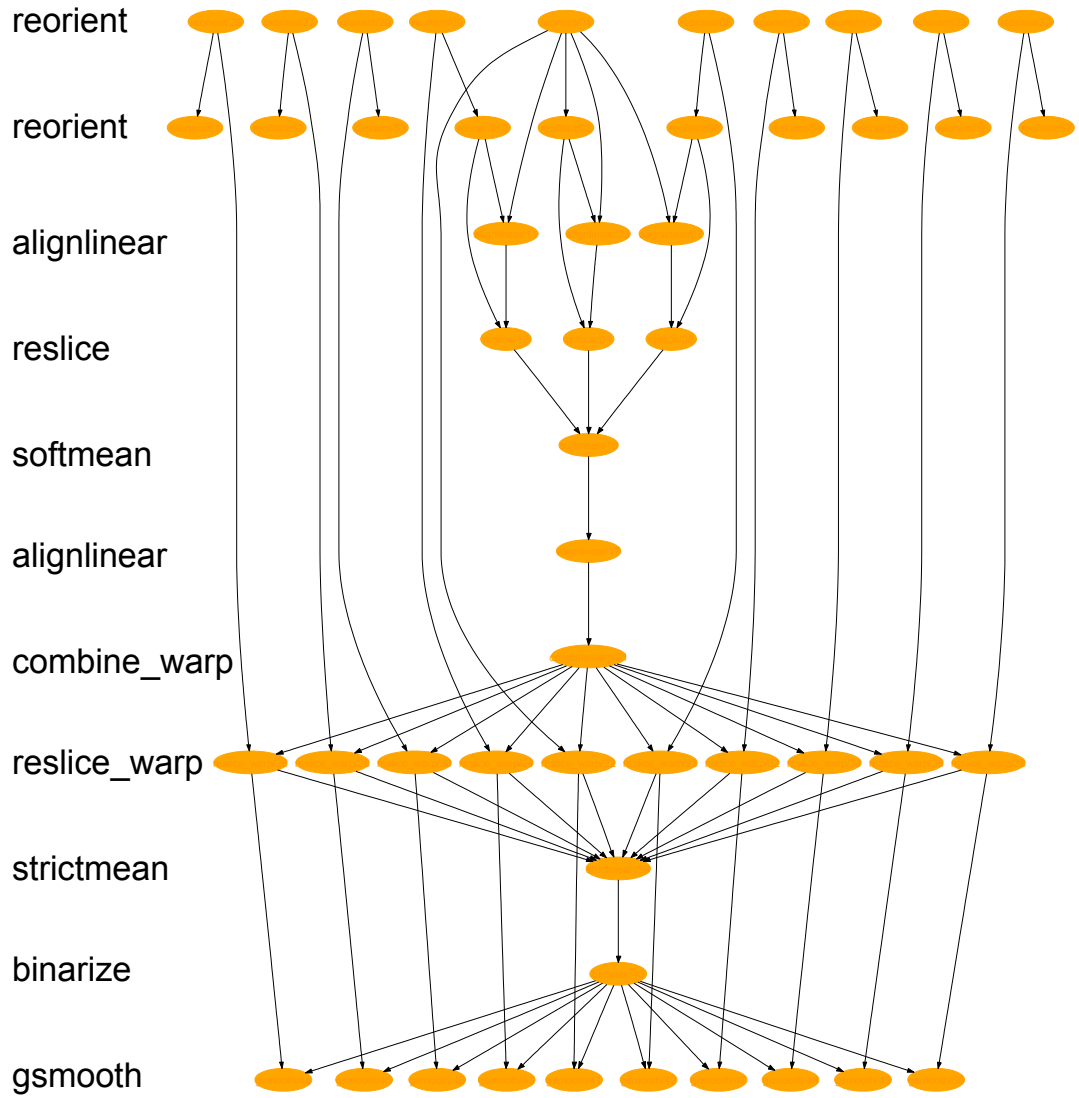
Work of: J Dennis, M Woitasek, S Mickelson, R Jacob



Spatial normalization of functional MRI runs



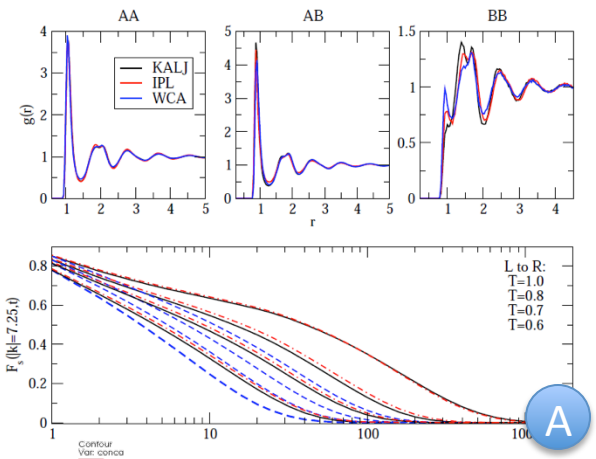
Dataset-level workflow

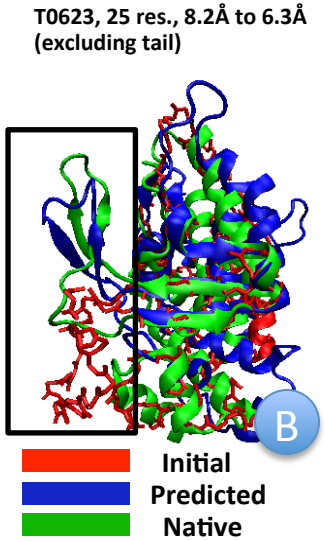


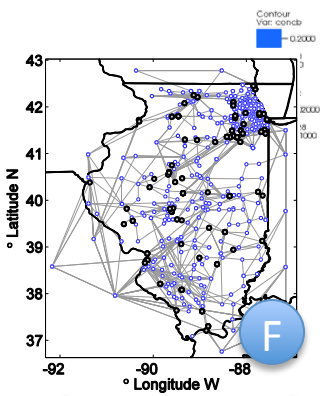
Expanded (10 volume) workflow

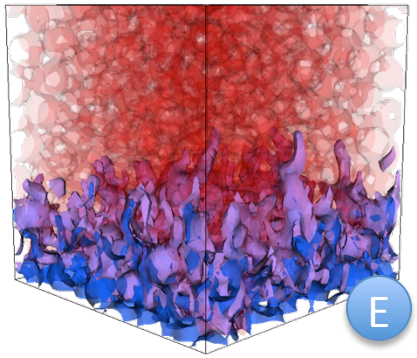
Numerous many-task workflow applications

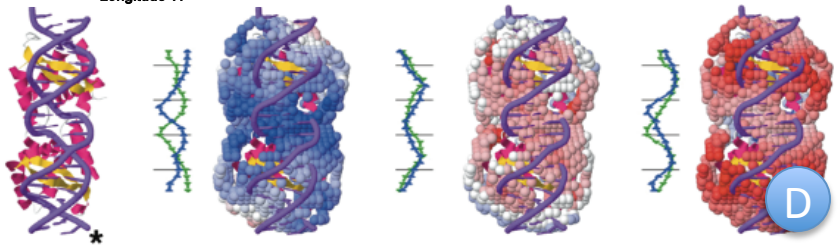
- **A** Simulation of super-cooled glass materials
- **B** Protein folding using homology-free approaches
- **C** Climate model analysis and decision making in energy policy
- **D** Simulation of RNA-protein interaction
- **E** Multiscale subsurface flow modeling
- **F** Modeling of power grid for OE applications

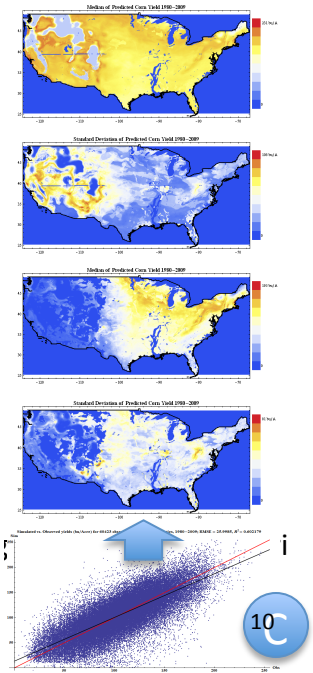








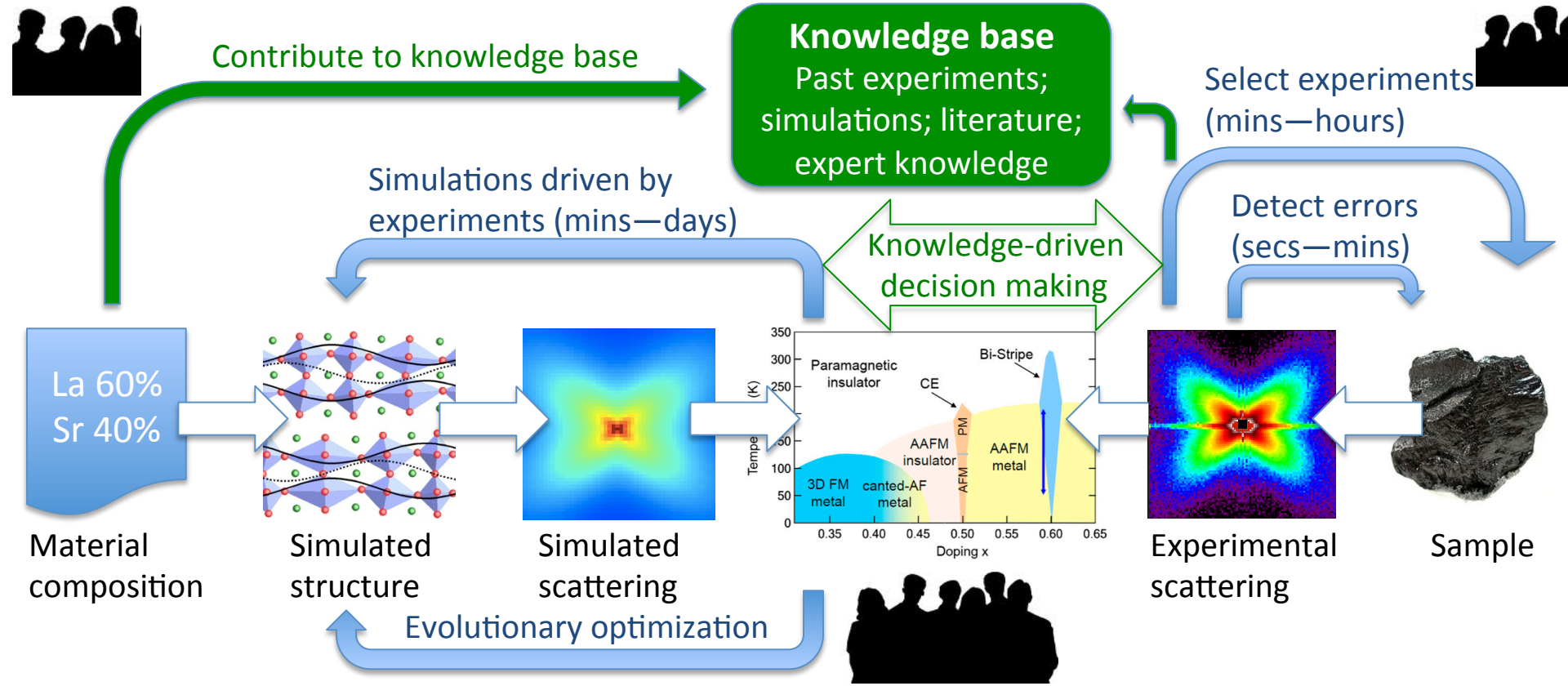




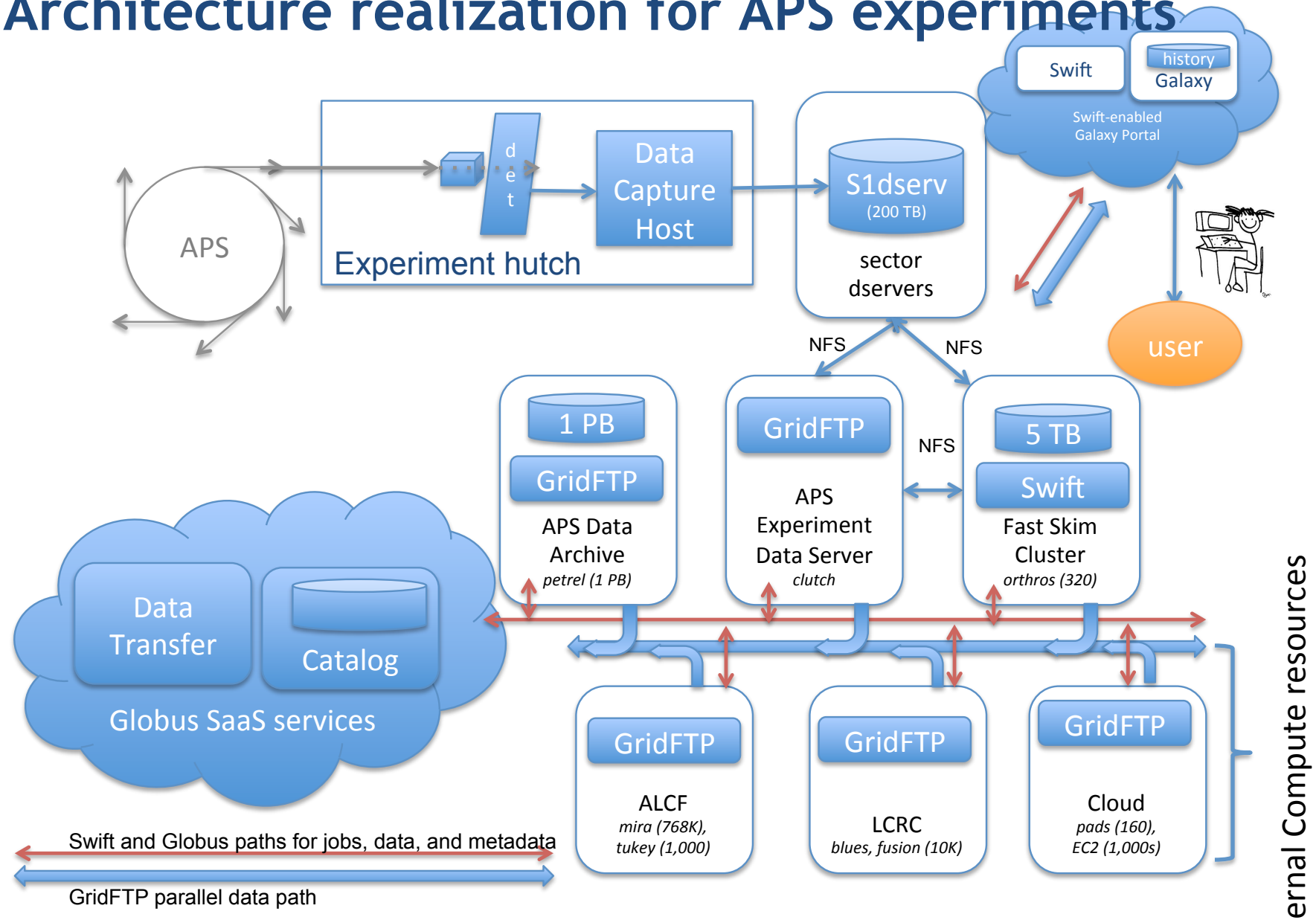
TARGET RESOURCES

- System types
  - Clouds
  - Clusters (campus, department)
  - Petascale HPC systems
  - Grids (OSG, LCG, …)
  - Multi/many-cores – 256 core nodes!
- Patterns
  - A single big HPC machine
  - HPC Machine with attached resources
  - Extend campus cluster with cloud
  - Many HPC machines
  - Many combinations of above

Diffuse scattering workflows



Architecture realization for APS experiments



Workflow patterns and issues

- Parameter sweeps
- Ensembles
- Data analysis
- Scaling studies
- Specialized patterns: uncertainty quantification, branch and bound
- Programming an application from libraries of applications
- Dataflow vs control flow
  - Ultimately, workflow is essentially dataflow
  - The difference is who writes and thinks about the dataflow
- Pipelining and concurrency (and how dataflow is good at this)
- Workflow manager drives application (outer workflow, inner scripts)
- Workflow manager embedded in application (outer scripts, inner workflow)

PROGRAMMING MODELS

- MPI, OpenMP, Hybrid
- Map reduce
- Record processing (with functions) vs file processing (with apps)
- Generating workflows for other engines
- Dynamically interpret the workflow
- Script mode (for Blue Gene, Cray systems)
- Dependent job processing

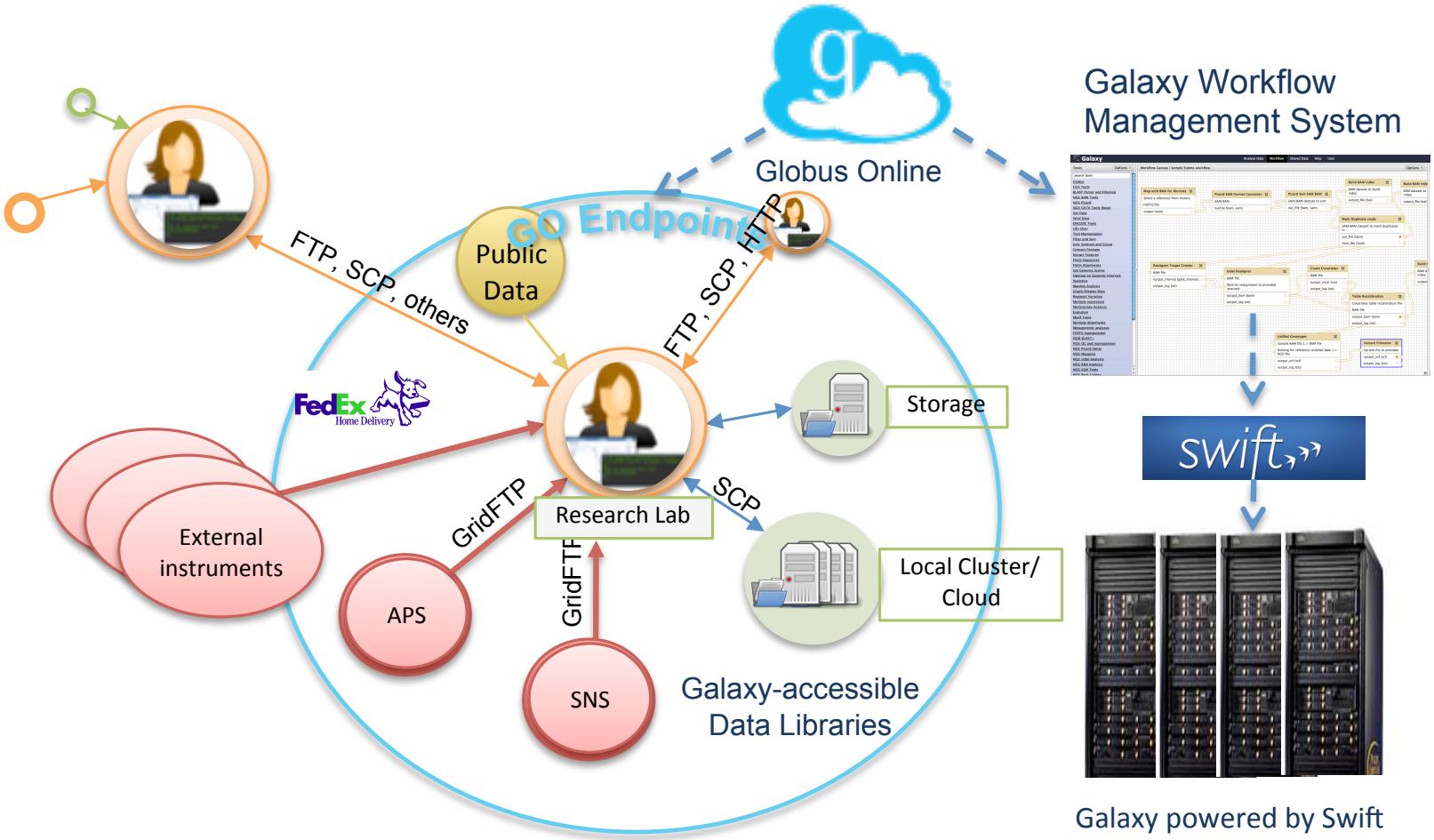
A *partial* sampler of workflow tools

- High throughput tools
  - Condor
  - Cluster schedulers / local resource managers (PBS, SGE, Cobalt, LSF, LL, SLURM,..)
- Workflow task dependency managers
  - DAGMan
  - Schedulers with job dependencies
- Integrated dependency and data management
  - Pegasus
- Dataflow languages
  - Dryad, Ciel, Swift
- Big data solutions
  - Hadoop, Spark, Zookeeper, Uzi
- Multicore tools
  - GNU Parallel, iPython parallel support
- Languages with parallel support
  - Py_nnn, Java_nnn, Haskell, R, MATLAB => PSOM, Parallel BASH (Walker)

A *sampler* of workflow tools (con't)

- Interactive workflow frameworks
  - Galaxy
  - Taverna
  - Kepler
  - LONI Pipeline (neuroscience)
  - Microsoft Workflow manager
  - Airivata
- Science gateways

Galaxy workflow portal



Data Management

Data Analysis

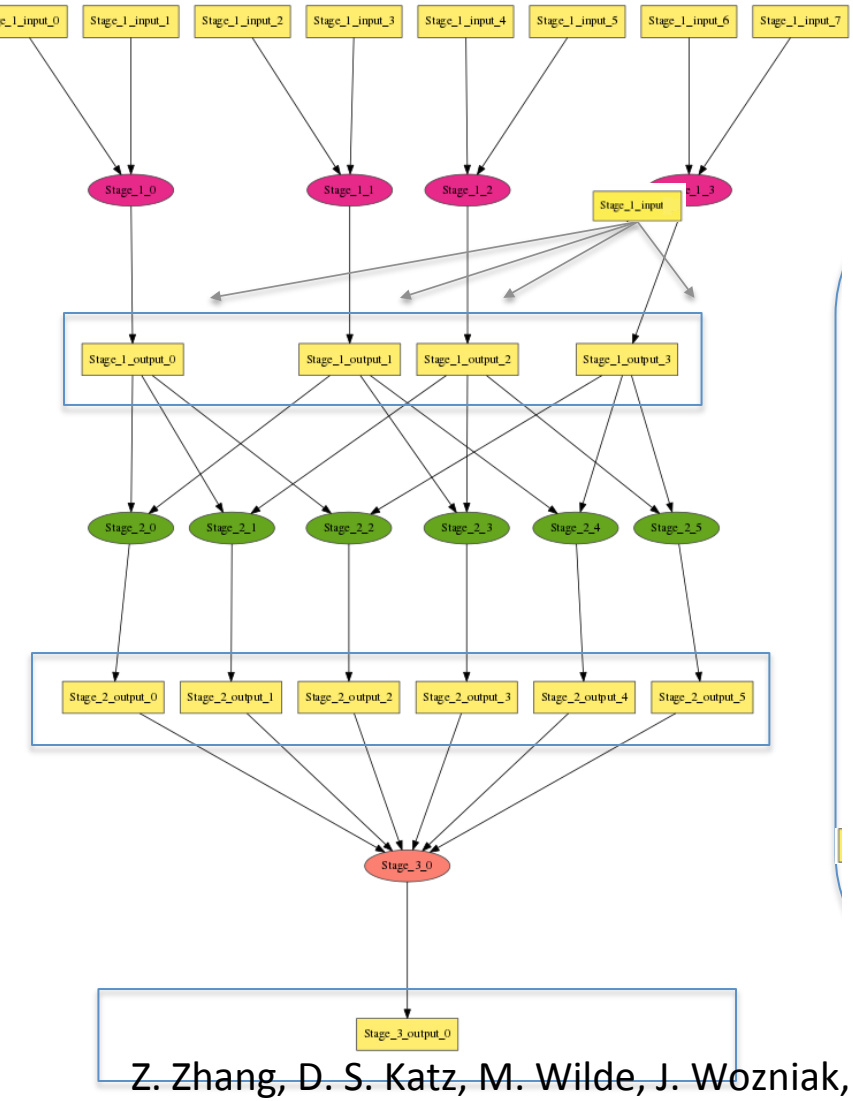
Two fundamental problems in scaling workflow

- Task rate
  - 60,000 cores / 60 sec/task = 1,000 tasks per second!
- Data management
  - 1K tasks / sec may generate 5GB/sec – not so bad if blocked efficiently
  - 1K tasks / sec may generate 2,000 files / sec – not so easy

Multi-level scheduling: pilot jobs can improve task rate performance

- Pilot jobs are long-running meta-jobs
  - allocate compute resources and run many smaller jobs
- PANDA
  - Widely used on OSG and LCG by the ATLAS physics collaboration
- GWMS using Condor Glide-Ins
  - A generalized solution widely deployed on OSG
- SAGA and Bigjob
  - Obtaining good results on XSEDE resources
- Java CoG Coasters
  - Allocates/frees resources based on demand
  - Peaks at 600 tasks per second
- Falkon
  - Research system reached 3,000 tasks per second and 1B tasks

Workflow patterns and data exchange



Filesystem Access Patterns:

- File Creation
- File Open
- 1-to-1 Read
- N-to-1 Read
- Few-to-1 Read
- 1-to-1 Write

Z. Zhang, D. S. Katz, M. Wilde, J. Wozniak, I. Foster. MTC Envelope: Defining the Capability of Large Scale Computers in the Context of Parallel Scripting Applications, HPDC 2013.

Some engineering problems and research challenges for extreme workflow

- Engineering
  - Diversity of interfaces, hard to tame and test, hard to abstract
  - Inter-language bindings and data interchange – challenge to usability
  - Integration with extreme-scale networks, runtimes and language stacks
- Research
  - Economics and policy-based scheduling
  - Retry/recovery of large distributed task and data graphs
  - Power management
  - Load balancing
  - Programming models: integration of dataflow and big-data techniques and tools

Summary: Challenges of workflow at extreme scale

- Inter-resource coordination
- Hybrid programming tools
- The challenges of data motion
  - Data management strategies and system envelopes
- The challenges of task scheduling and dispatch
  - Task rates and task distribution
  - Resource utilization vs. time to solution
- Workflow expression and separation of concerns
- Provenance: tracking what was done

The Swift parallel scripting language

Presenter: Michael Wilde wilde@mcs.anl.gov

Mathematics and Computer Science Division
Argonne National Laboratory
University of Chicago/Argonne Computation Institute

swift

- Parallel scripting language for clusters, clouds & grids
  - For writing loosely-coupled scripts of application programs and utilities linked by exchanging files
  - Can call scripts in shell, python, R, Octave, MATLAB, …
- Swift does 3 important things for you:
  - Makes parallelism transparent – with functional dataflow
  - Makes basic failure recovery transparent
  - *Makes computing location transparent* – can run your script on multiple distributed sites and diverse computing resources (from desktop to petascale)

Swift programming model: all progress driven by concurrent dataflow

```
(int r) myproc (int i, int j)
{
    int f = F(i);
    int g = G(j);
    r = f + g;
}
```

- `F()` and `G()` implemented in native code or external programs
- `F()` and `G()` run in concurrently in different processes
- `r` is computed when they are both done

- This parallelism is *automatic*
- Works recursively throughout the program's call graph

Swift programming model

- Data types

```
int    i = 4;
int    A[];
string s = "hello world";
```

- Mapped data types

```
file image<"snapshot.jpg">;
```

- Structured data

```
image  A[]<array_mapper…>;
type protein {
   file pdb;
   file docking_pocket;
}
protein p<ext; exec=protein.map>;
```

- Conventional expressions

```
if (x == 3) {
    y = x+2;
    s = @strcat("y: ", y);
}
```

- Parallel loops

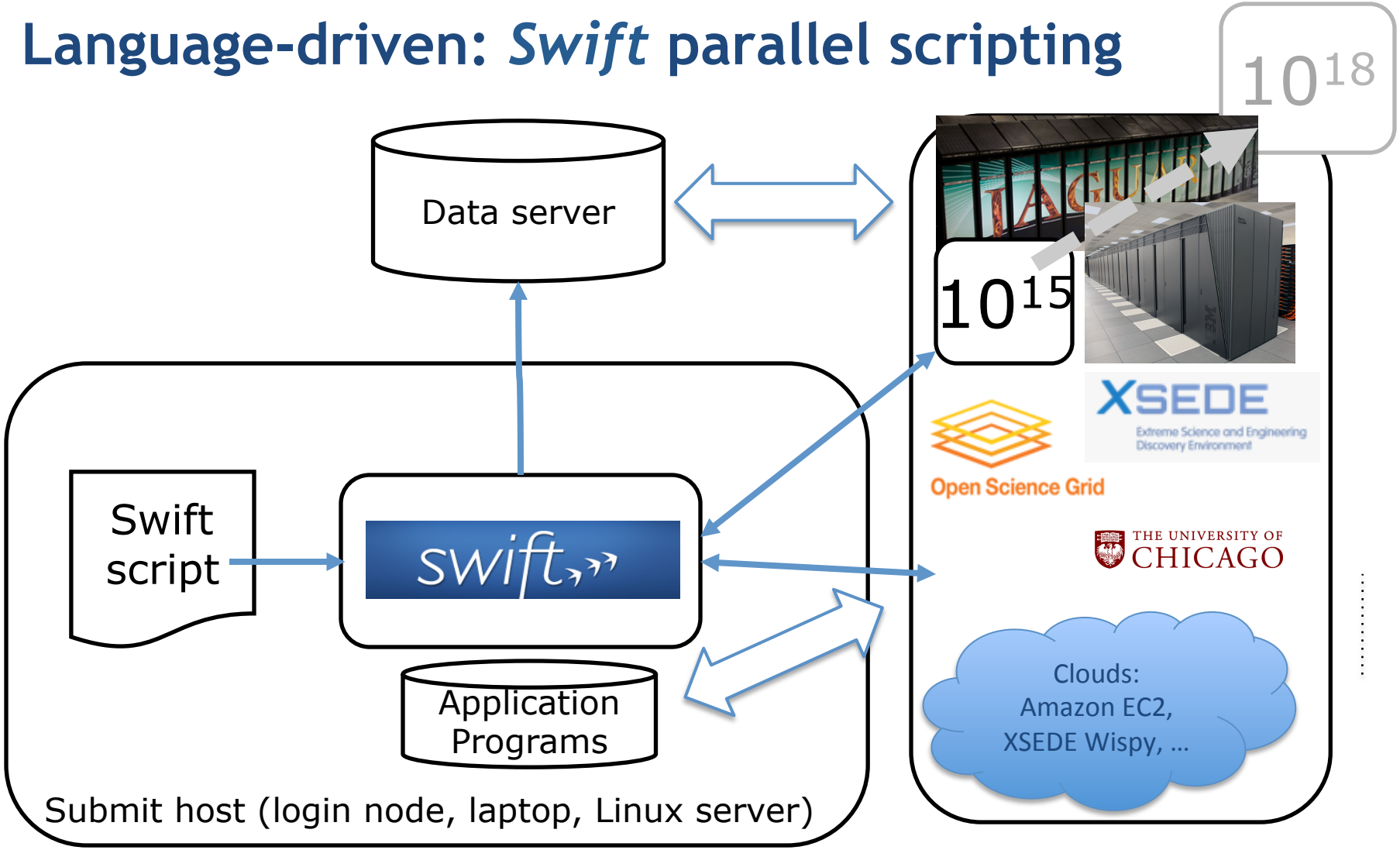
```
foreach f,i in A {
    B[i] = convert(A[i]);
}
```

- Data flow

```
analyze(B[0], B[1]);
analyze(B[2], B[3]);
```

Swift: A language for distributed parallel scripting, J. Parallel Computing, 2011

Language-driven: *Swift* parallel scripting



Swift runs parallel scripts on a broad range of parallel computing resources.

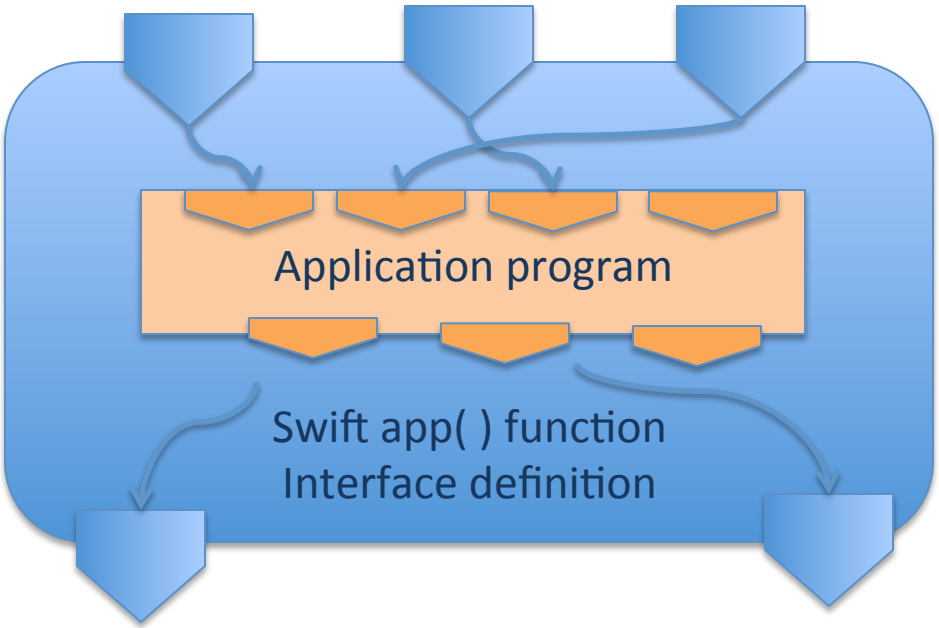
Programming model: all execution driven by parallel data flow

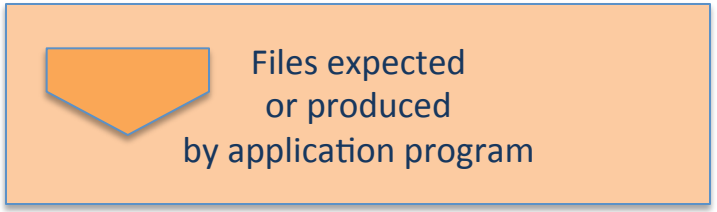
```
(int r) myproc (int i)
{
    j = f(i);
    k = g(i);
    r = j + k;
}
```

- f() and g() are computed in parallel
- myproc() returns r when they are done

- This parallelism is *automatic*
- Works recursively throughout the program's call graph

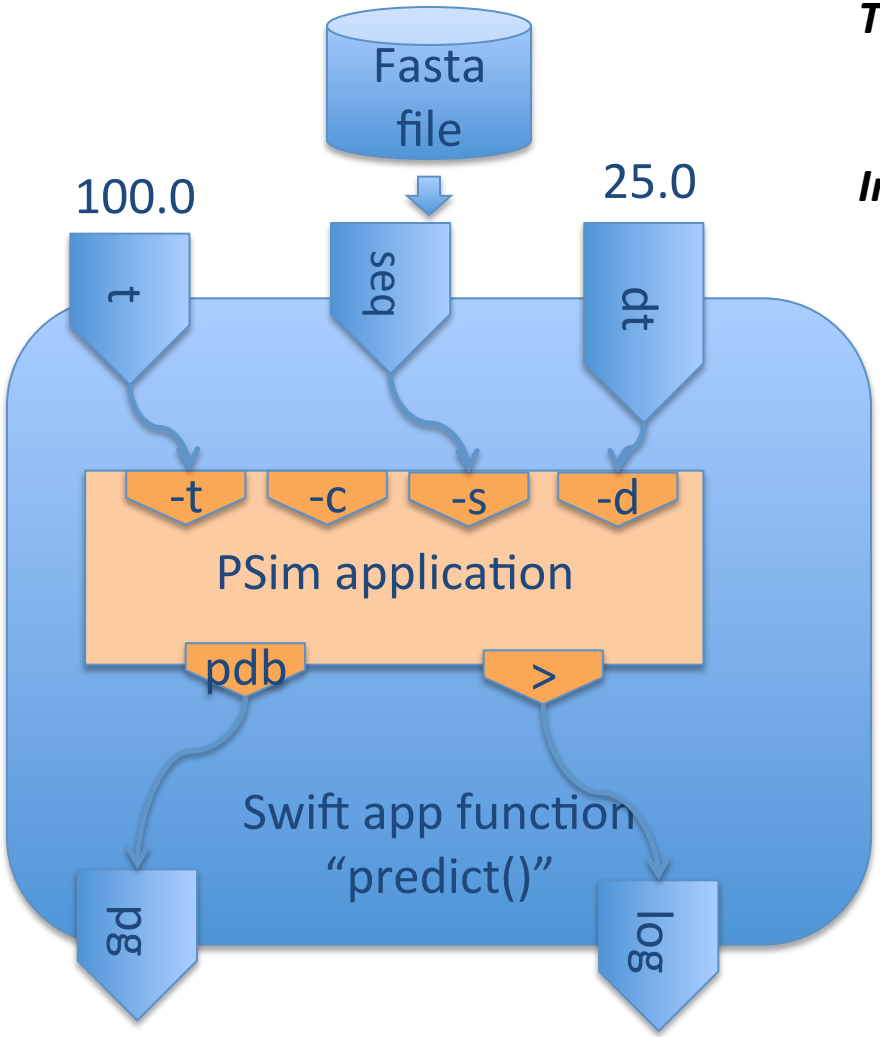
Encapsulation enables distributed parallelism





Encapsulation is the key to transparent distribution, parallelization, and automatic provenance capture

app( ) functions specify cmd line argument passing



To run:

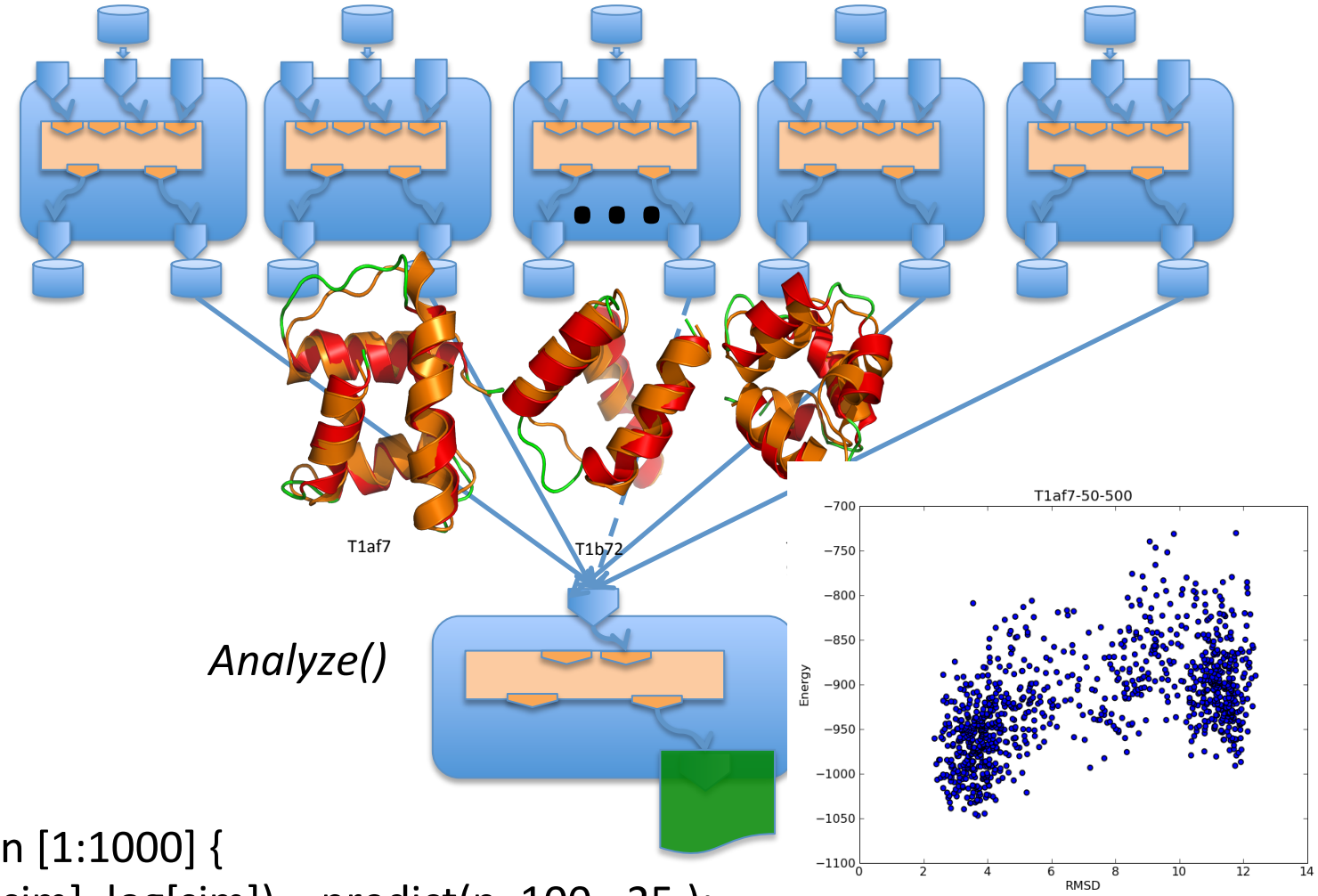
psim -s 1ubq.fas -pdb p -t 100.0 -d 25.0 >log

In Swift code:

```
app (PDB pg, File log) predict (Protein seq,
                                Float t, Float dt)
{
  psim "-c" "-s" @pseq.fasta "-pdb" @pg
       "–t" temp "-d" dt;
}

Protein p <ext; exec="Pmap", id="1ubq">;
PDB structure;
File log;

(structure, log) = predict(p, 100., 25.);
```

Large scale parallelization with simple loops

1000 Runs of the "predict" application



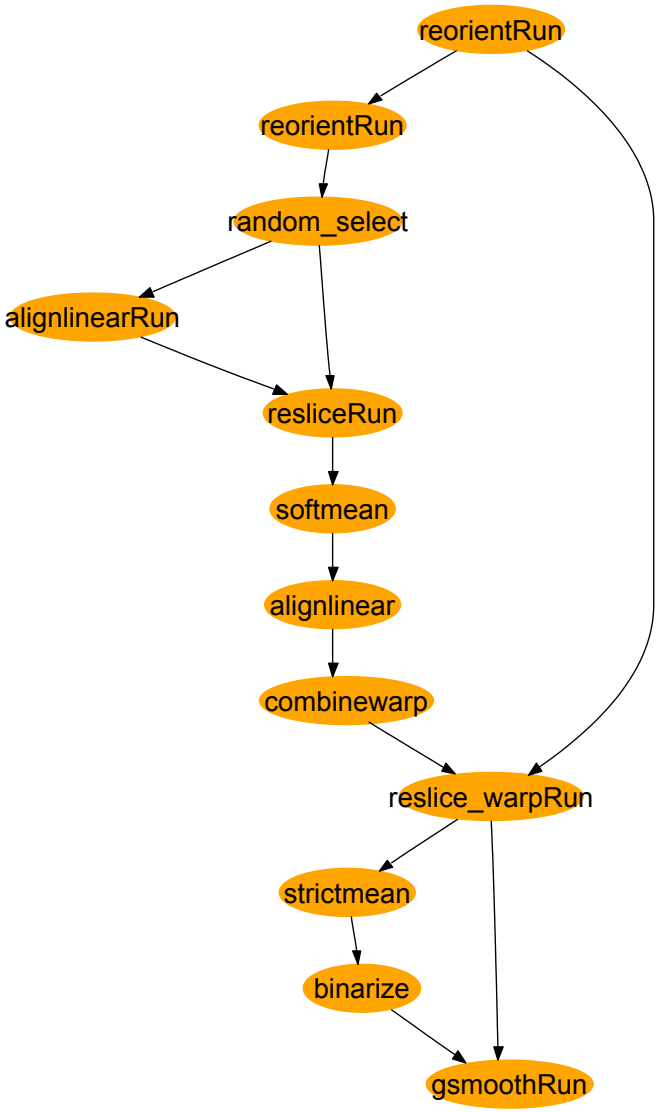

```
foreach sim in [1:1000] {
   (structure[sim], log[sim]) = predict(p, 100., 25.);
}
result = analyze(structure)
```

Nested parallel prediction loops in *Swift*

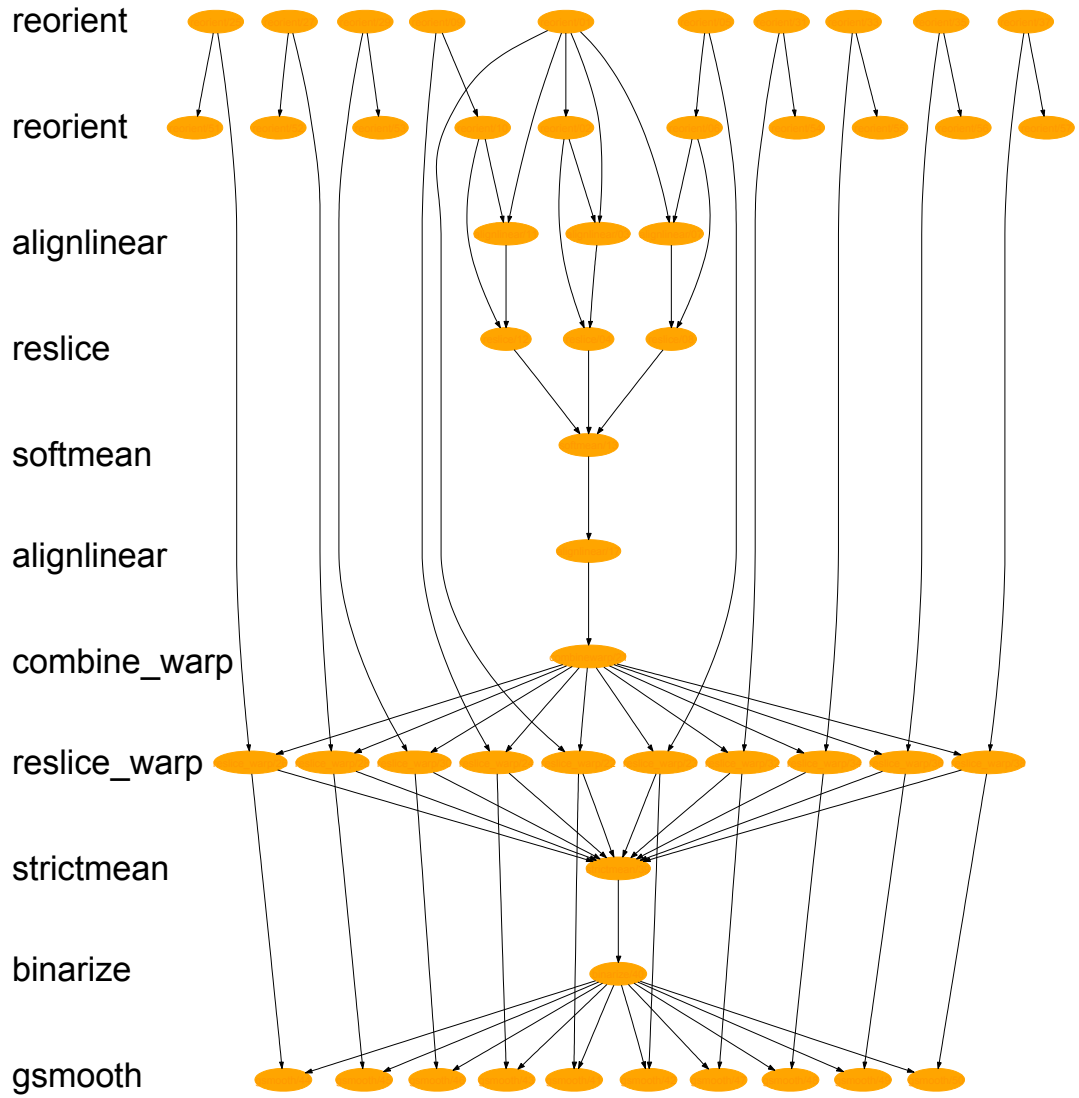
```
Sweep( )
{
  int nSim = 1000;
  int maxRounds = 3;
  Protein pSet[ ] <ext; exec="Protein.map">;
  float startTemp[ ] = [ 100.0, 200.0 ];
  float delT[ ] = [ 1.0, 1.5, 2.0, 5.0, 10.0 ];
  foreach p, pn in pSet {
    foreach t in startTemp {
      foreach d in delT {
        ItFix(p, nSim, maxRounds, t, d);
      }
    }
  }
}
Sweep();
```

10 proteins x 1000 simulations x
3 rounds x 2 temps x 5 deltas
= 300K tasks

Spatial normalization of functional run



Dataset-level workflow



Expanded (10 volume) workflow

Complex scripts can be well-structured

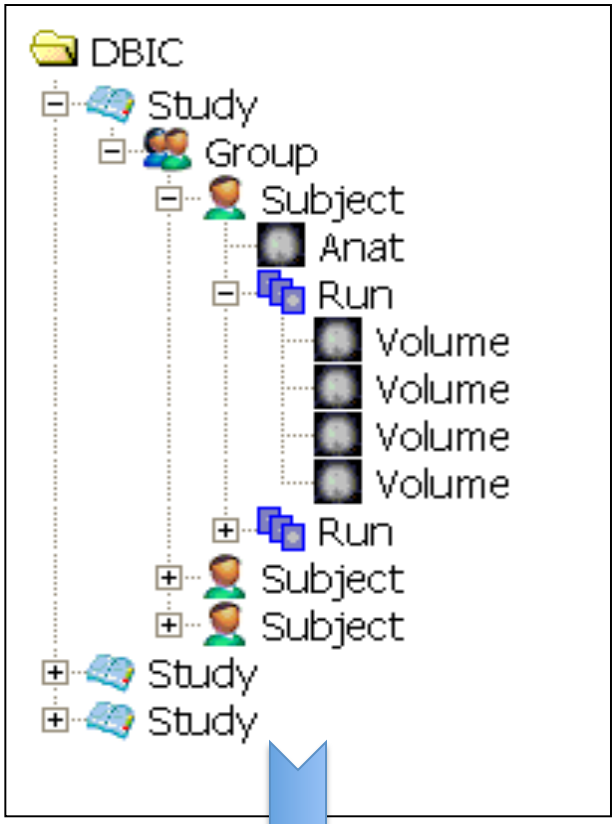
programming in the large: fMRI spatial normalization script example

```
(Run snr) functional ( Run r, NormAnat a,
                       Air shrink )
{    Run yroRun = reorientRun( r , "y" );
     Run roRun = reorientRun( yroRun , "x" );
     Volume std = roRun[0];
     Run rndr = random_select( roRun, 0.1 );
     AirVector rndAirVec = align_linearRun( rndr, std, 12, 1000, 1000, "81 3 3" );
     Run reslicedRndr = resliceRun( rndr, rndAirVec, "o", "k" );
     Volume meanRand = softmean( reslicedRndr, "y", "null" );
     Air mnQAAir = alignlinear( a.nHires, meanRand, 6, 1000, 4, "81 3 3" );
     Warp boldNormWarp = combinewarp( shrink, a.aWarp, mnQAAir );
     Run nr = reslice_warp_run( boldNormWarp, roRun );
     Volume meanAll = strictmean( nr, "y", "null" )
     Volume boldMask = binarize( meanAll, "y" );
     snr = gsmoothRun( nr, boldMask, "6 6 6" );
}
```

```
(Run or) reorientRun ( Run ir, string direction)
{
        foreach Volume iv, i in ir.v {
                or.v[i] = reorient(iv, direction);
        }
}
```

Dataset mapping example: fMRI datasets

On-Disk Data Layout



Mapping function or script

```
type Study {
        Group g[ ];
}

type Group {
        Subject s[ ];
}

type Subject {
        Volume anat;
        Run run[ ];
}

type Run {
        Volume v[ ];
}

type Volume {
        Image img;
        Header hdr;
}
```

Swift's in-memory data model

Nested loops can generate massive parallelism

Protein folding example:

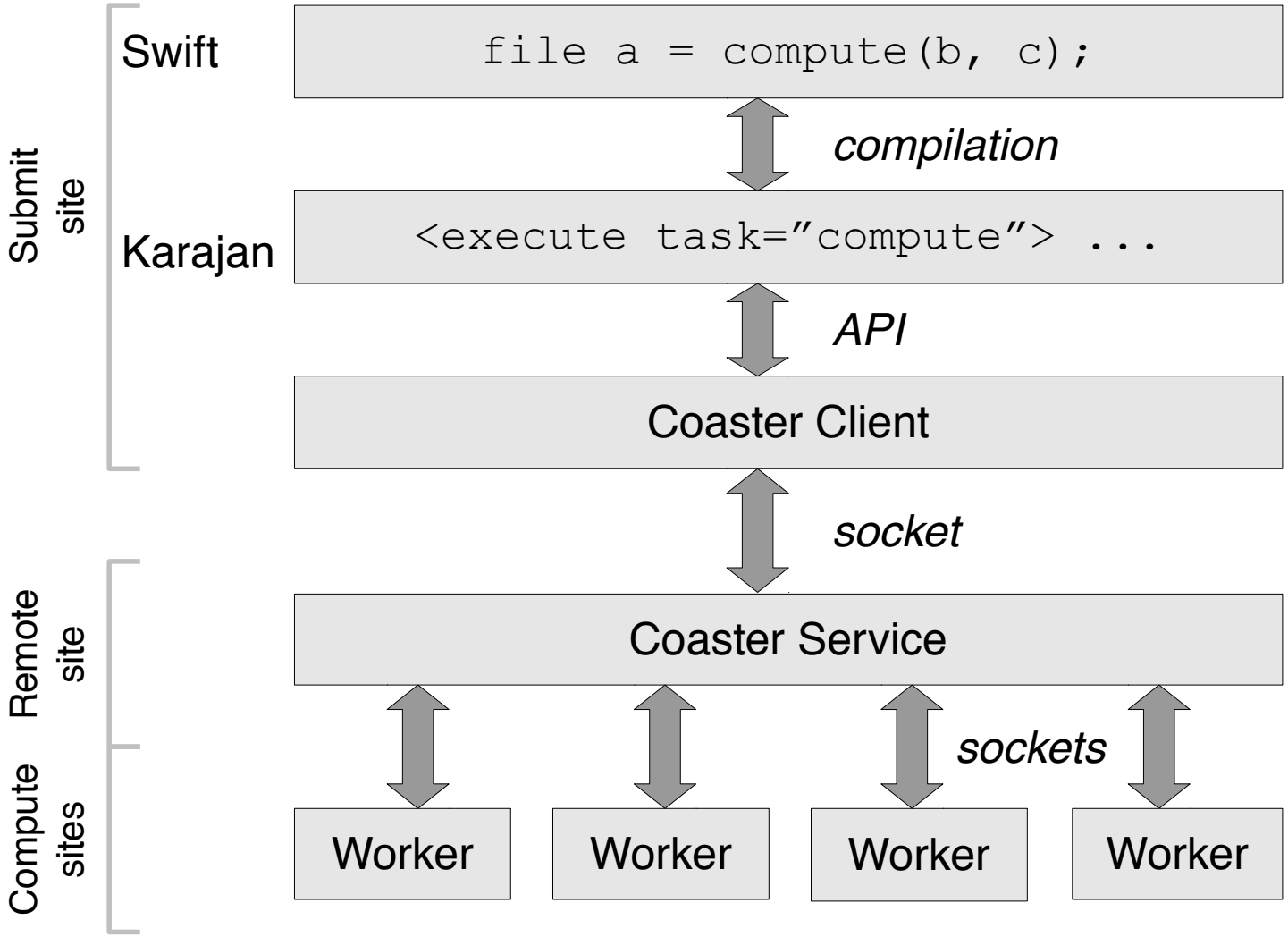
```
Sweep( )
{
   int nSim = 1000;
   int maxRounds = 3;
   Protein pSet[ ] <ext; exec="Protein.map">;
   float startTemp[ ] = [ 100.0, 200.0 ];
   float delT[ ] = [ 1.0, 1.5, 2.0, 5.0, 10.0 ];
   foreach p, pn in pSet {
      foreach t in startTemp {
         foreach d in delT {
            ItFix(p, nSim, maxRounds, t, d);
         }
      }
   }
}
Sweep();
```

10 proteins x 1000 simulations x
3 rounds x 2 temps x 5 deltas
= 300K tasks

Flexible worker-node agents for execution and data transport

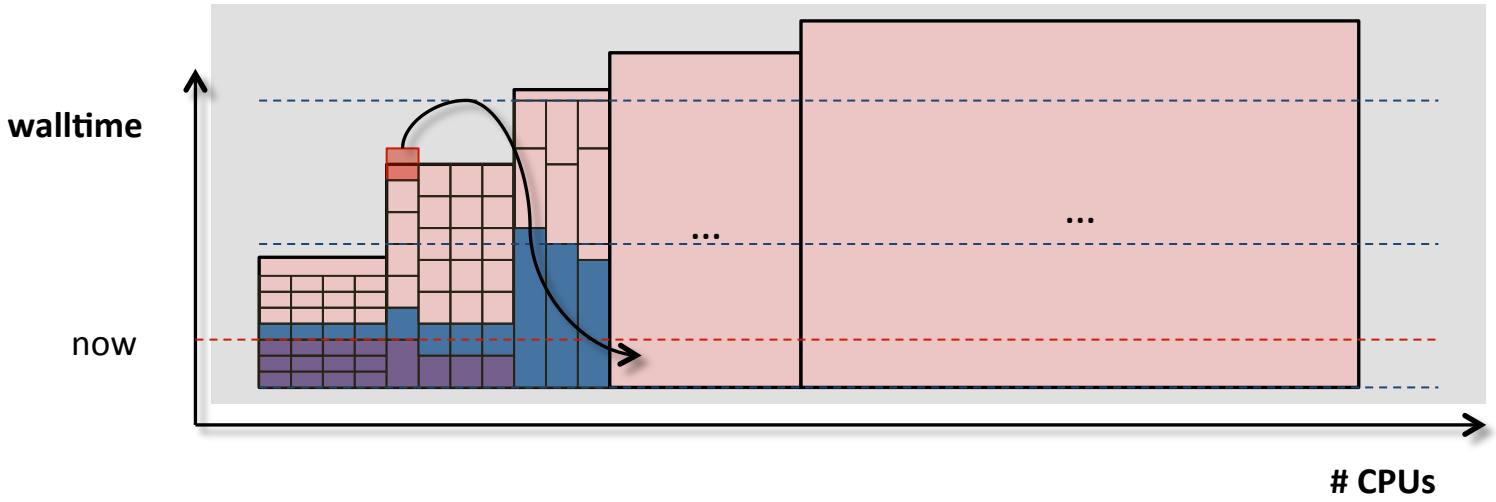
- Main role is to efficiently run Swift tasks on allocated compute nodes, local and remote
- Handles short tasks efficiently
- Runs over Grid infrastructure: Condor, GRAM
- Also runs with few infrastructure dependencies
- Can optionally perform data staging
- Provisioning can be automatic or external (manually launched or externally scripted)

Worker architecture handles diverse environments



Implementation: The job packing problem (II) (also not to scale)

- Commit jobs to blocks and adjust as necessary based on actual walltime



- The actual packing problem is NP-complete
- Solved using a greedy algorithm: always pick the largest job that will fit in a block first

- Swift is a parallel scripting system for grids, clouds and clusters
  - for loosely-coupled applications - application and utility programs linked by exchanging files
- Swift is easy to write: simple high-level C-like functional language
  - Small Swift scripts can do large-scale work
- Swift is easy to run: contains all services for running Grid workflow - in one Java application
  - Untar and run – acts as a self-contained Grid client
- Swift is fast: uses efficient, scalable and flexible "Karajan" execution engine.
  - Scaling close to 1M tasks – .5M in live science work, and growing
- Swift usage is growing:
  - applications in neuroscience, proteomics, molecular dynamics, biochemistry, economics, statistics, and more.
- Try Swift! http://swift-lang.org (Swift/K) and www.mcs.anl.gov/exm (Swift/T)

Swift/T
High Performance Workflow Language

Justin M Wozniak

wozniak@mcs.anl.gov

Goal: Programmability for large scale analysis

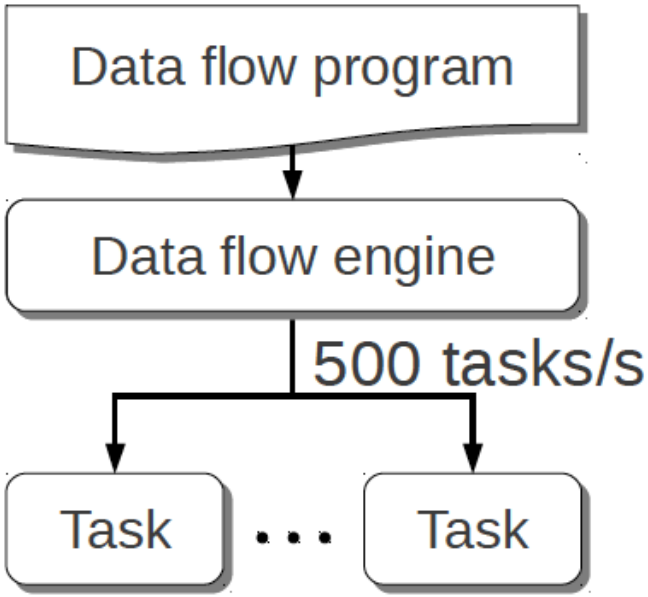
- Our solution is "many-task" computing: higher-level applications composed of many run-to-completion tasks: input→compute→output
  Message passing is handled by our implementation details

- Programmability
  - Large number of applications have this natural structure at upper levels: Parameter studies, ensembles, Monte Carlo, branch-and-bound, stochastic programming, UQ
  - Coupling extreme-scale applications to preprocessing, analysis, and visualization
- Data-driven computing
  - Dataflow-based execution models
  - Data organization tools in the programming languages
- Challenges
  - Load balancing, data movement, expressibility

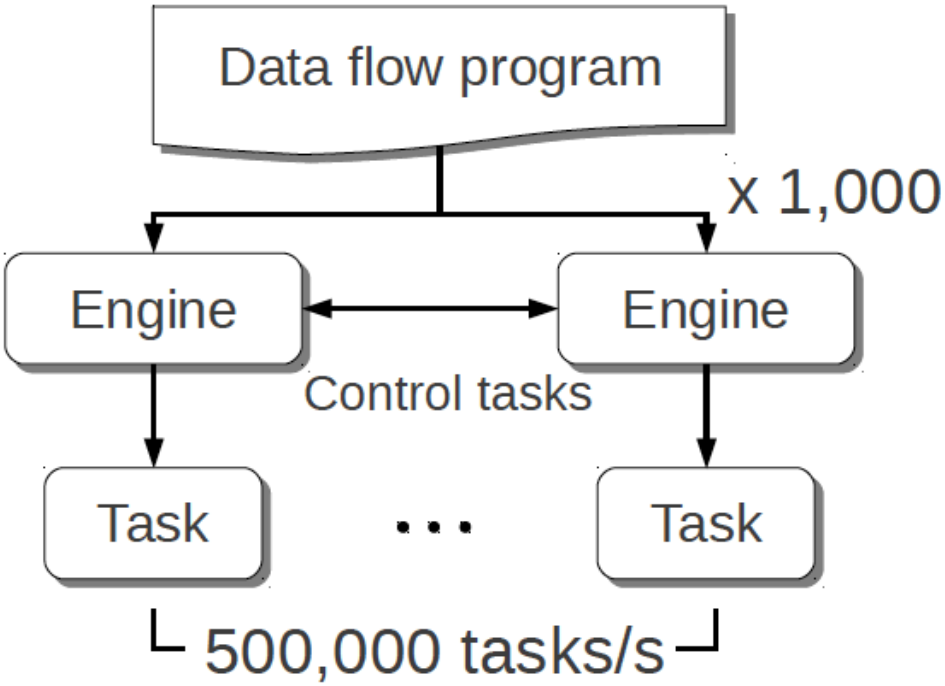
SWIFT/T OVERVIEW

Swift/T: Swift for high-performance computing

Had this:
(Swift/K)



For extreme scale, we need this:
(Swift/T)

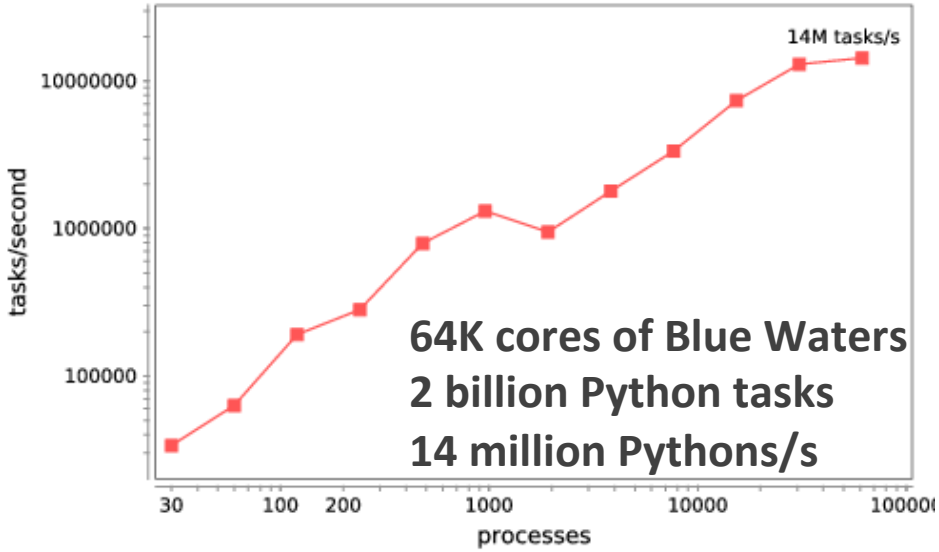


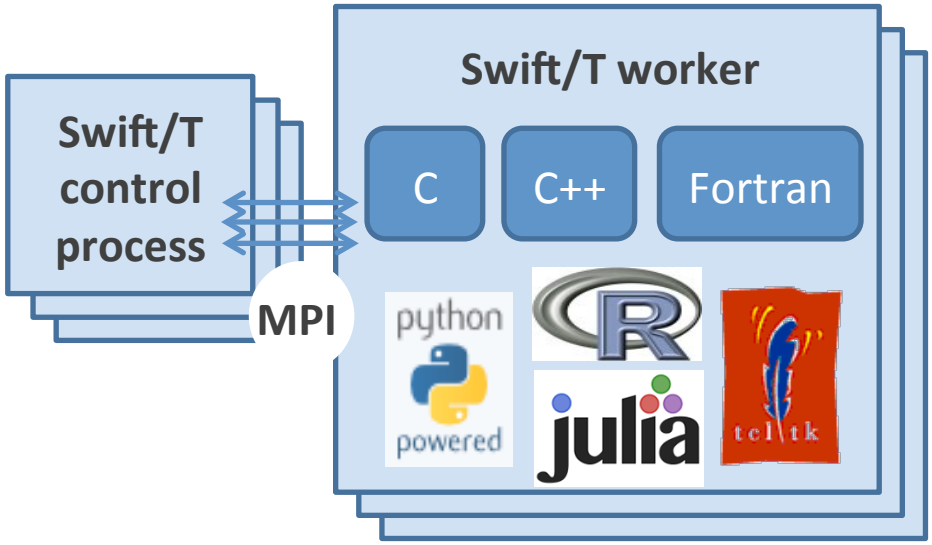
Centralized evaluation

Distributed evaluation

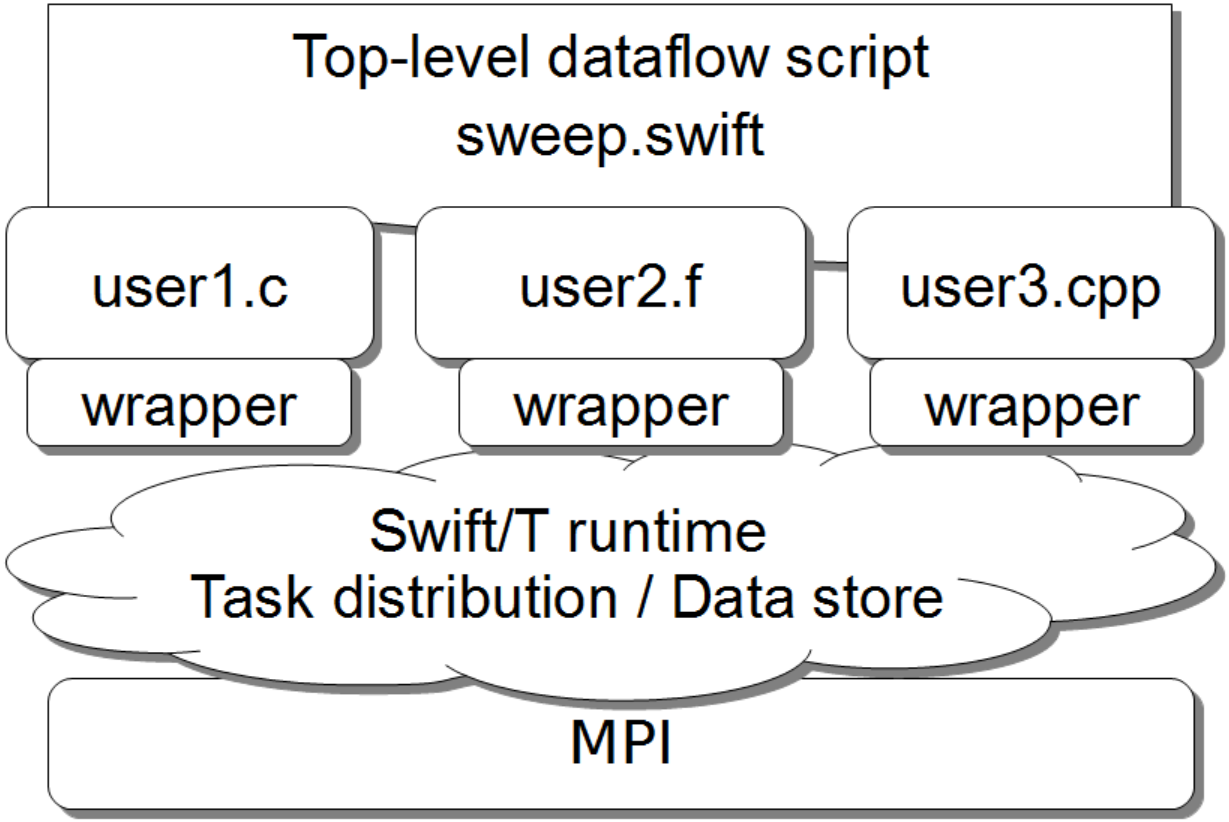
Swift/T: Enabling high-performance workflows

- Write site-independent scripts
- Automatic parallelization and data movement
- Run native code, script fragments as applications
- Rapidly subdivide large partitions for MPI jobs
- **Move work to data locations**





Support calls to native libraries



- Including MPI libraries

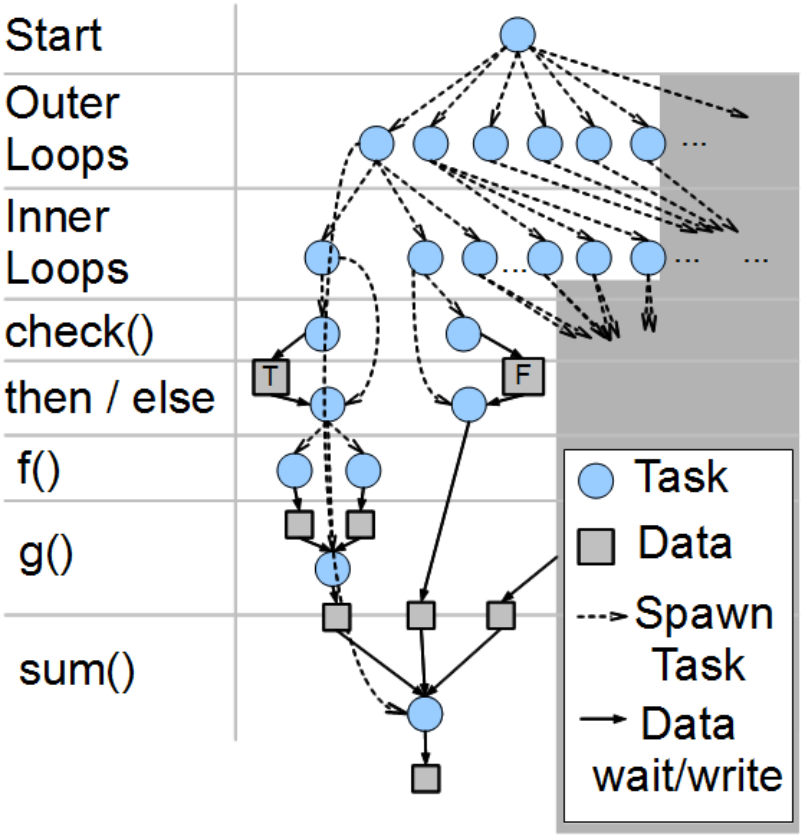
Characteristics of very large Swift programs

```
int X = 100, Y = 100;
int A[][];
int B[];
foreach x in [0:X-1] {
  foreach y in [0:Y-1] {
    if (check(x, y)) {
      A[x][y] = g(f(x), f(y));
    } else {
      A[x][y] = 0;
    }
  }
  B[x] = sum(A[x]);
}
```

- The goal is to support billion-way concurrency: $O(10^9)$
- Swift script logic will control trillions of variables and data dependent tasks
- Need to distribute Swift logic processing over the HPC compute system

Swift/T: Fully parallel evaluation of complex scripts

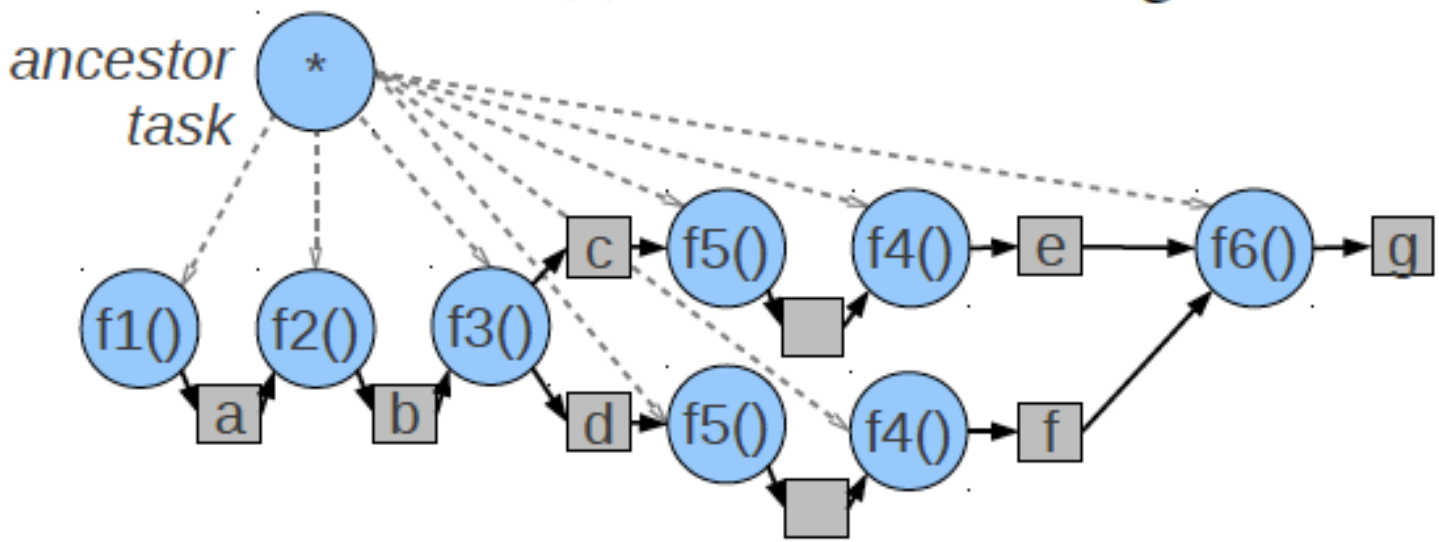
```
int X = 100, Y = 100;
int A[][];
int B[];
foreach x in [0:X-1] {
  foreach y in [0:Y-1] {
    if (check(x, y)) {
      A[x][y] = g(f(x), f(y));
    } else {
      A[x][y] = 0;
    }
  }
  B[x] = sum(A[x]);
}
```



Swift/T optimization challenge: distributed vars

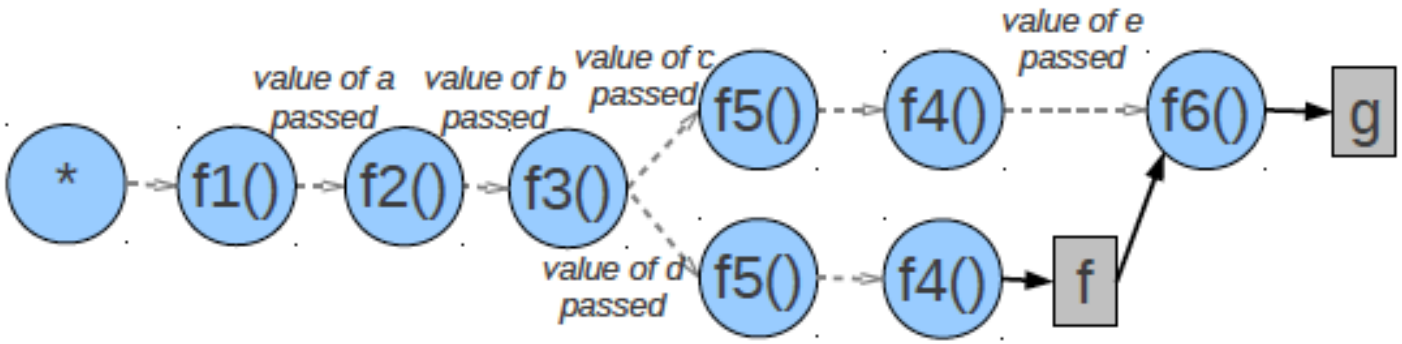
```
1 | a = f1();            b = f2(a);
2 | c, d = f3(a, b);     e = f4(f5(c);
3 | f = f4(f5(d);        g = f6(e, f);
```

(a) Swift/T code fragment

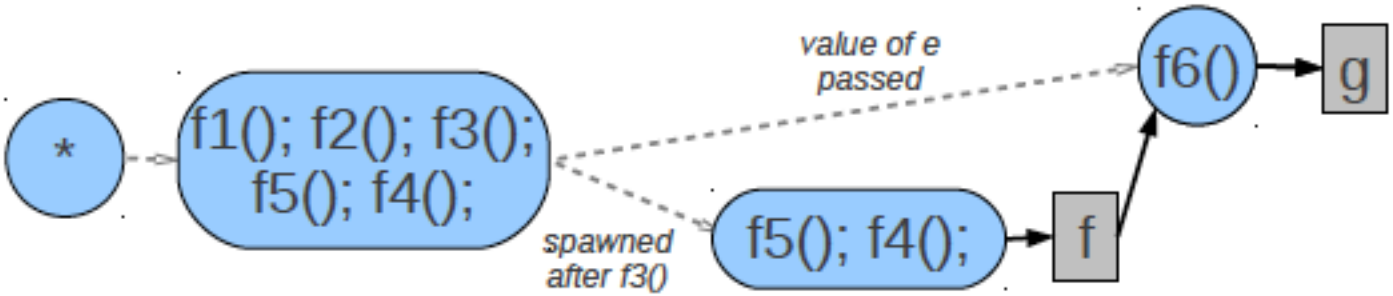


(b) Unoptimized version, passing data as shared data and perform synchronization

Swift/T optimizations improve data locality



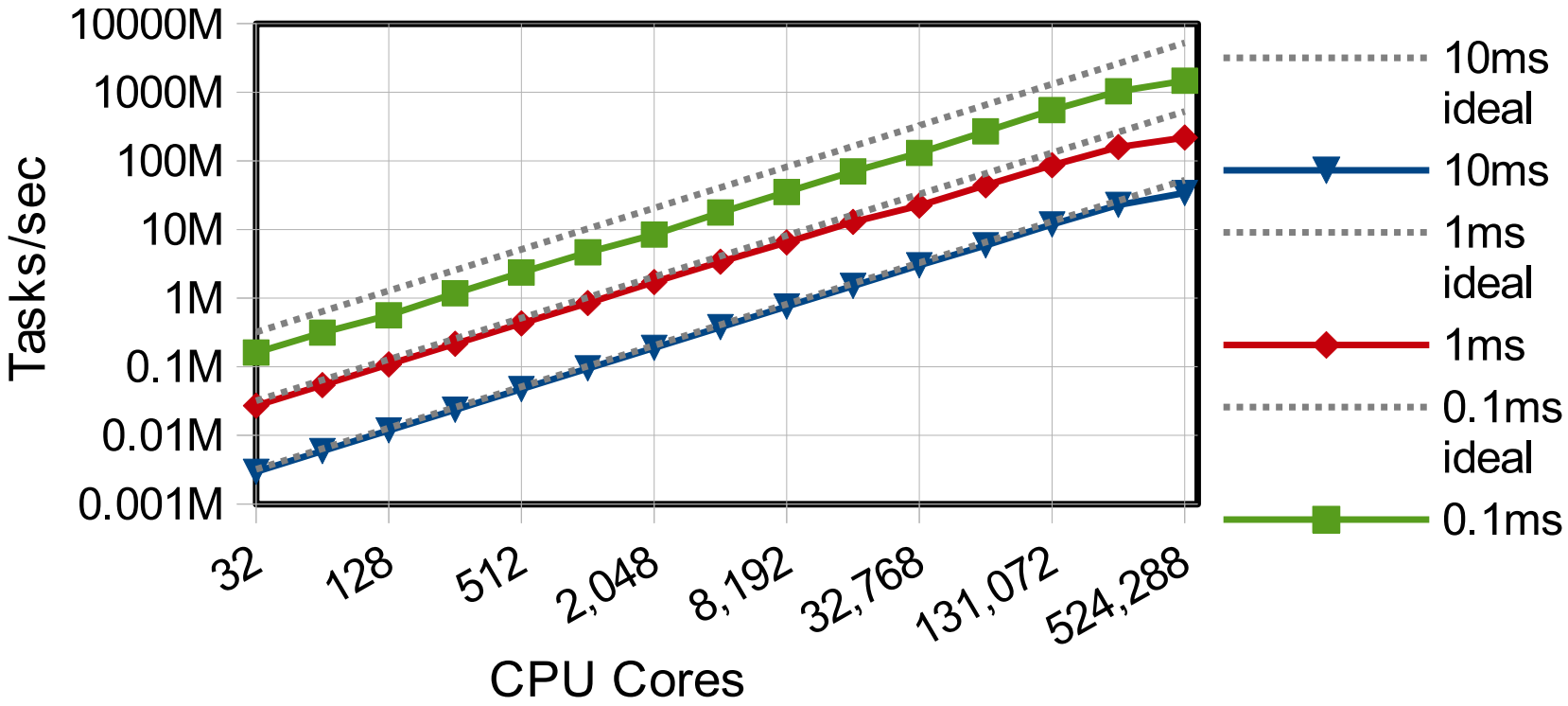
(c) After wait pushdown and elimination of shared data in favor of parent-to-child data passing



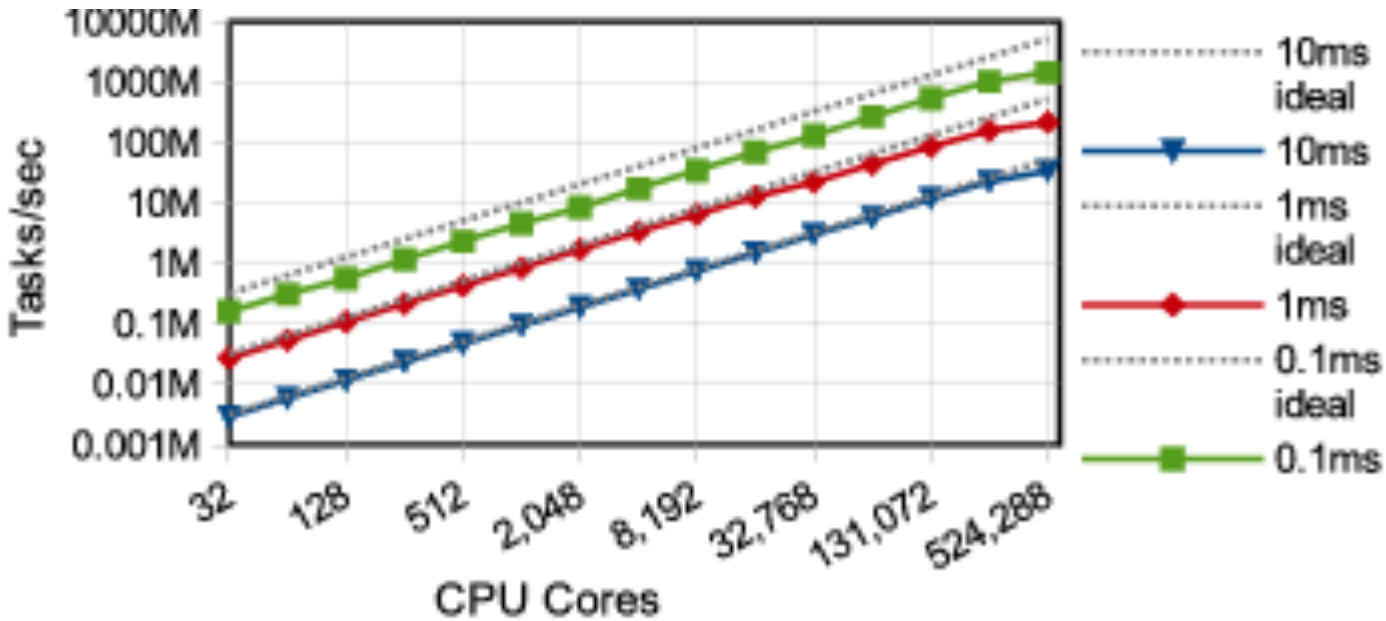
(d) After pipeline fusion merges tasks

Swift/T: scaling of trivial foreach { } loop

100 microsecond to 10 millisecond tasks on up to 512K integer cores of Blue Waters



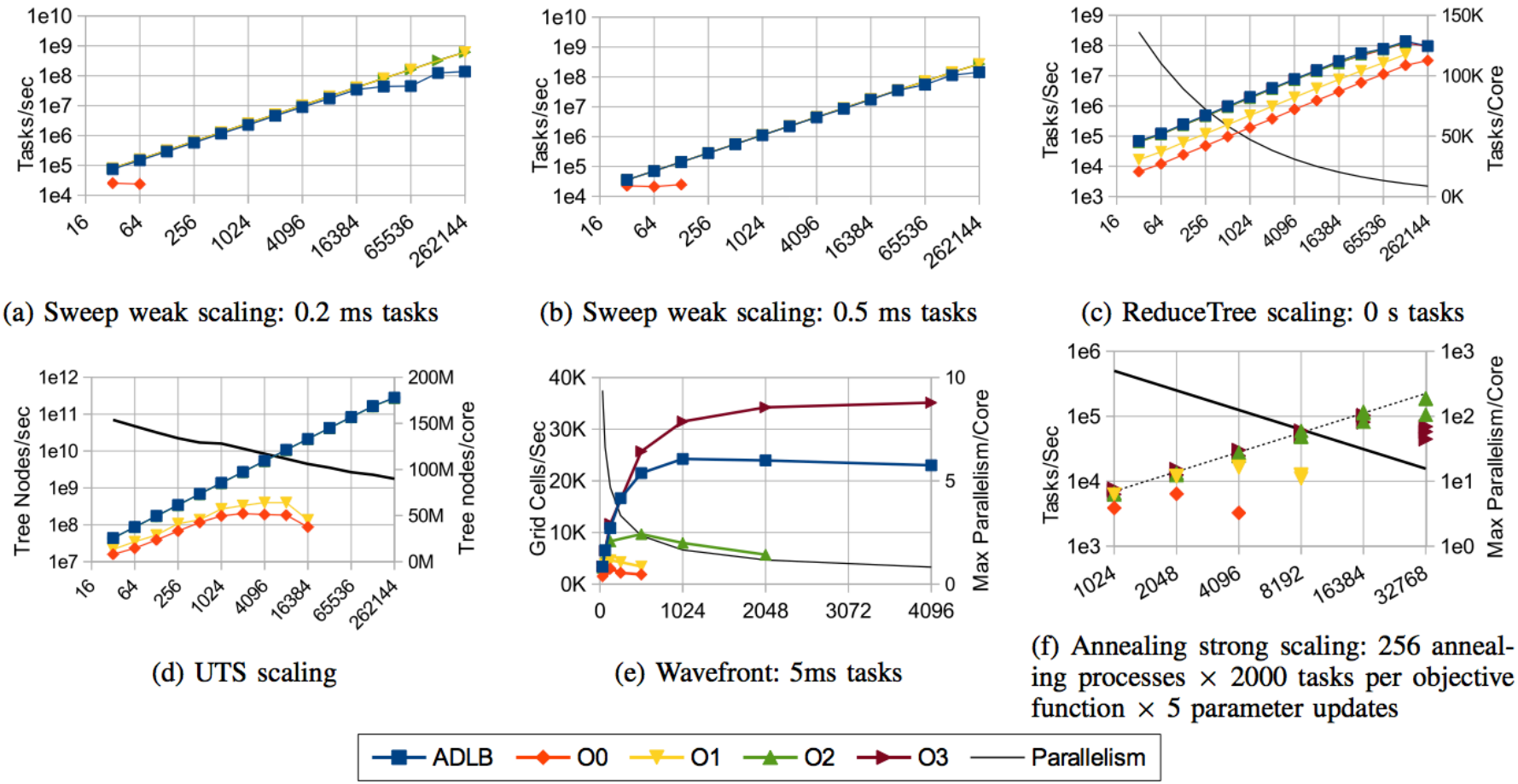
Basic scalability



- 1.5 billion tasks/s on 512K cores of Blue Waters, so far

- Armstrong et al. Compiler techniques for massively scalable implicit task parallelism. Proc. SC 2014.

Swift/T application benchmarks on Blue Waters



(a) Sweep weak scaling: 0.2 ms tasks

(b) Sweep weak scaling: 0.5 ms tasks

(c) ReduceTree scaling: 0 s tasks

(d) UTS scaling

(e) Wavefront: 5ms tasks

(f) Annealing strong scaling: 256 annealing processes × 2000 tasks per objective function × 5 parameter updates

Fig. 10: Application speedup and scalability at different optimization levels. X axes show scale in cores. Primary Y axes show application throughput in application-dependent terms. Secondary Y axes show problem size or degree of parallelism where applicable.

GeMTC: GPU-enabled Many-Task Computing

Motivation: Support for MTC on all accelerators!

Goals:

1) MTC support 2) Programmability

3) Efficiency 4) MPMD on SIMD

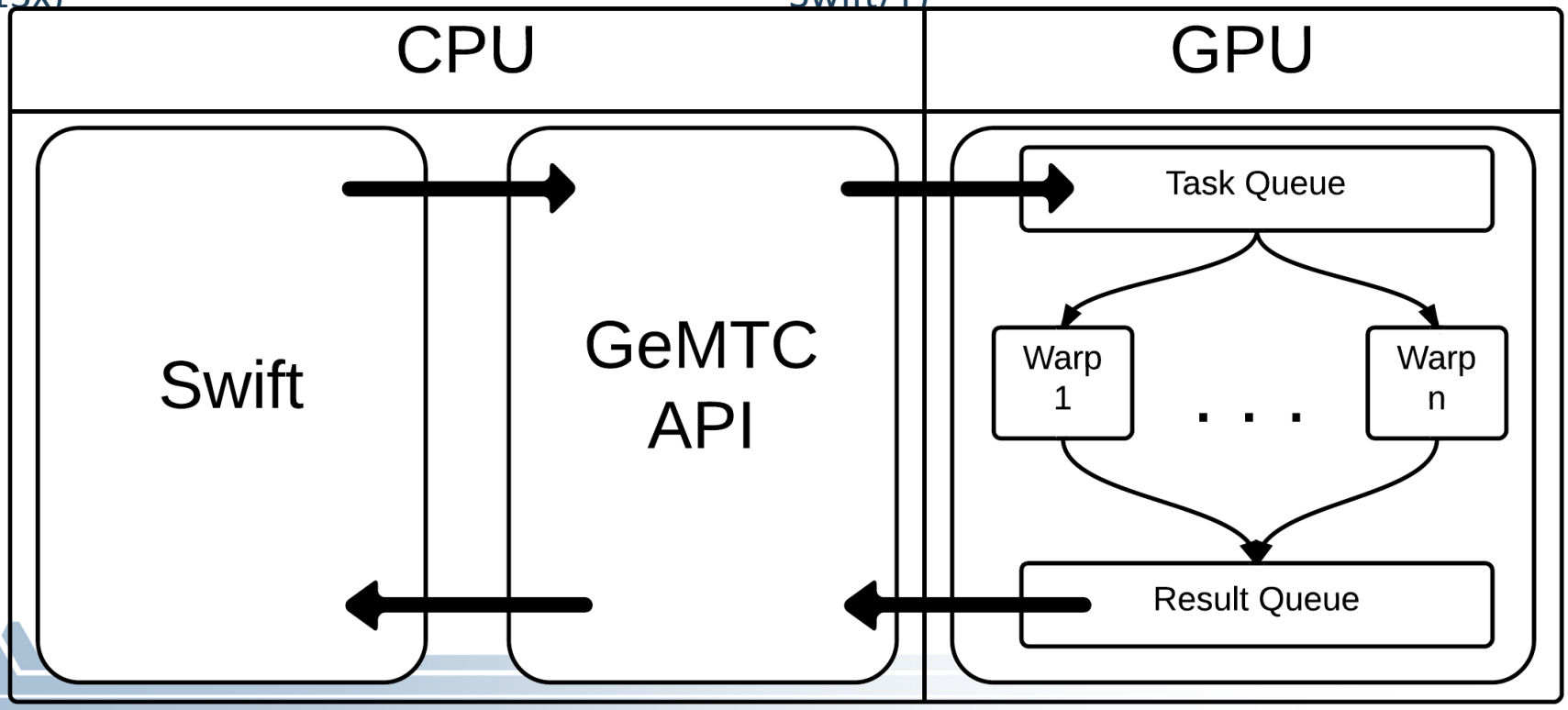
5) Increase concurrency from 15 to 192 (~13x)

Approach:

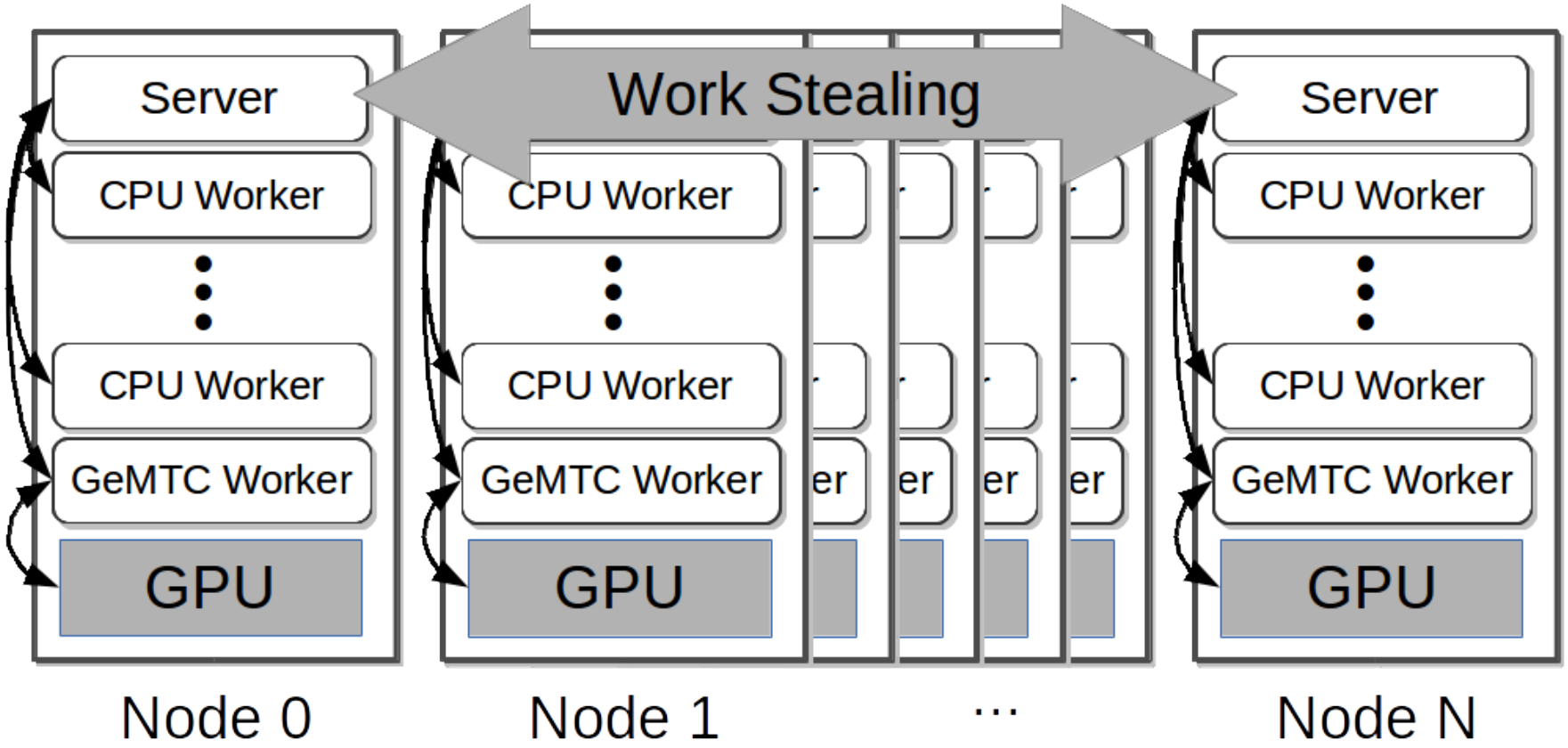
Design & implement GeMTC middleware:

1) Manages GPU 2) Spread host/device

3) Workflow system integration (with Swift/T)

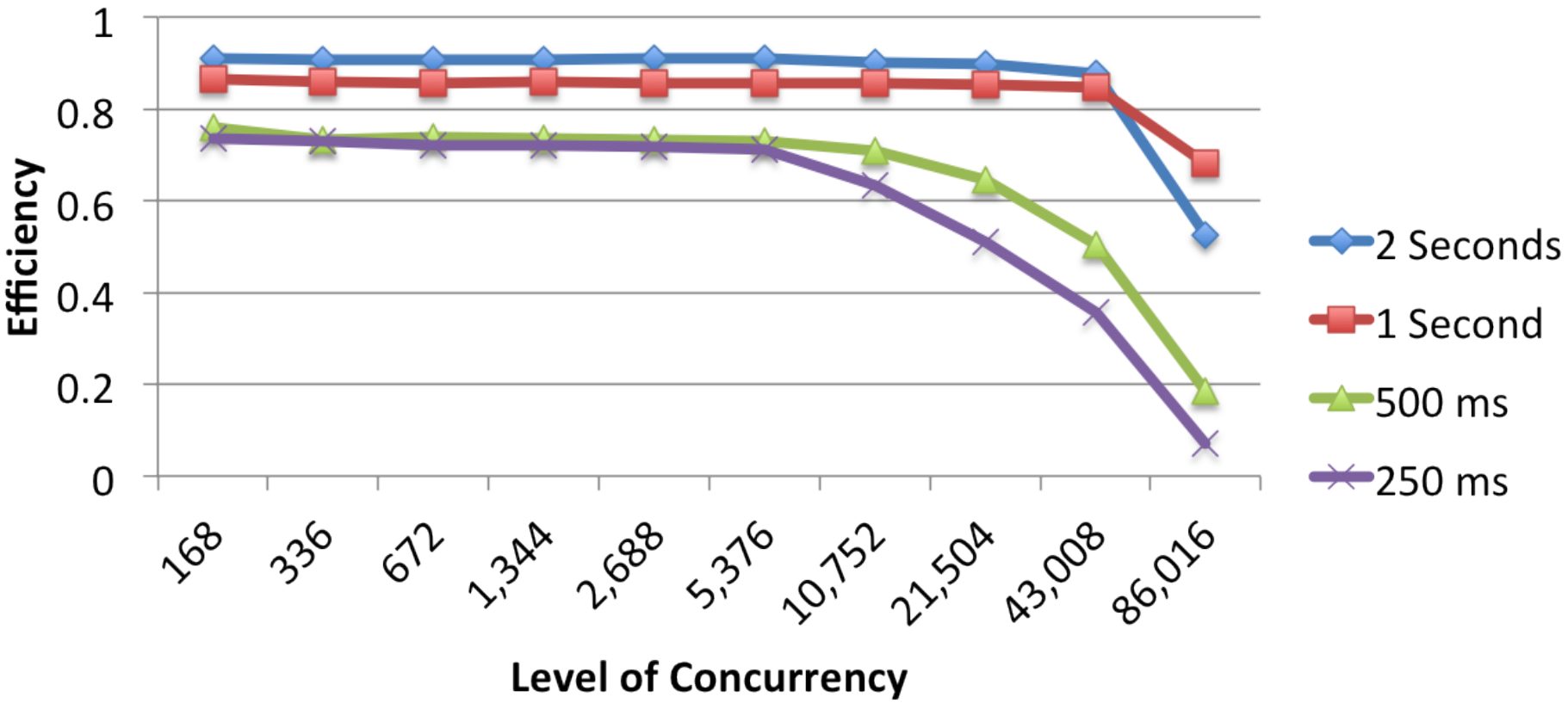


Swift/T + GeMTC Node Layout



GeMTC Efficiency: 512 Nodes, 86K GPU Workers

168 active workers per GPU



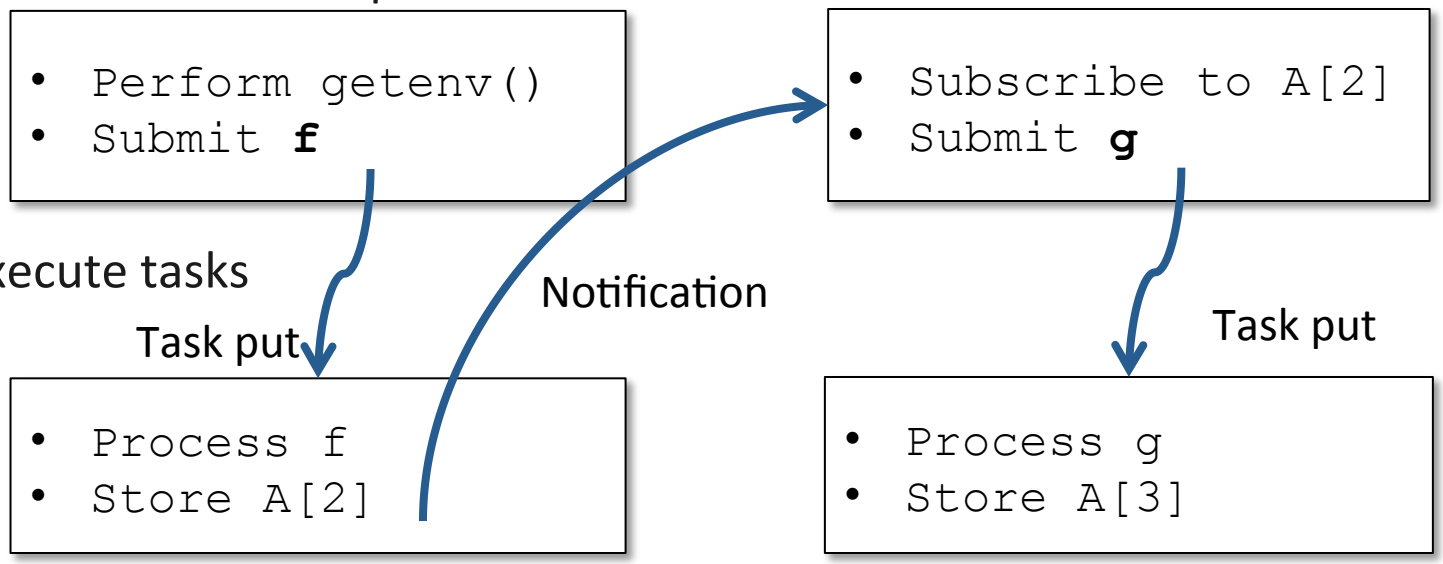
Example execution

- Code

```
A[2] = f(getenv("N"));
A[3] = g(A[2]);
```

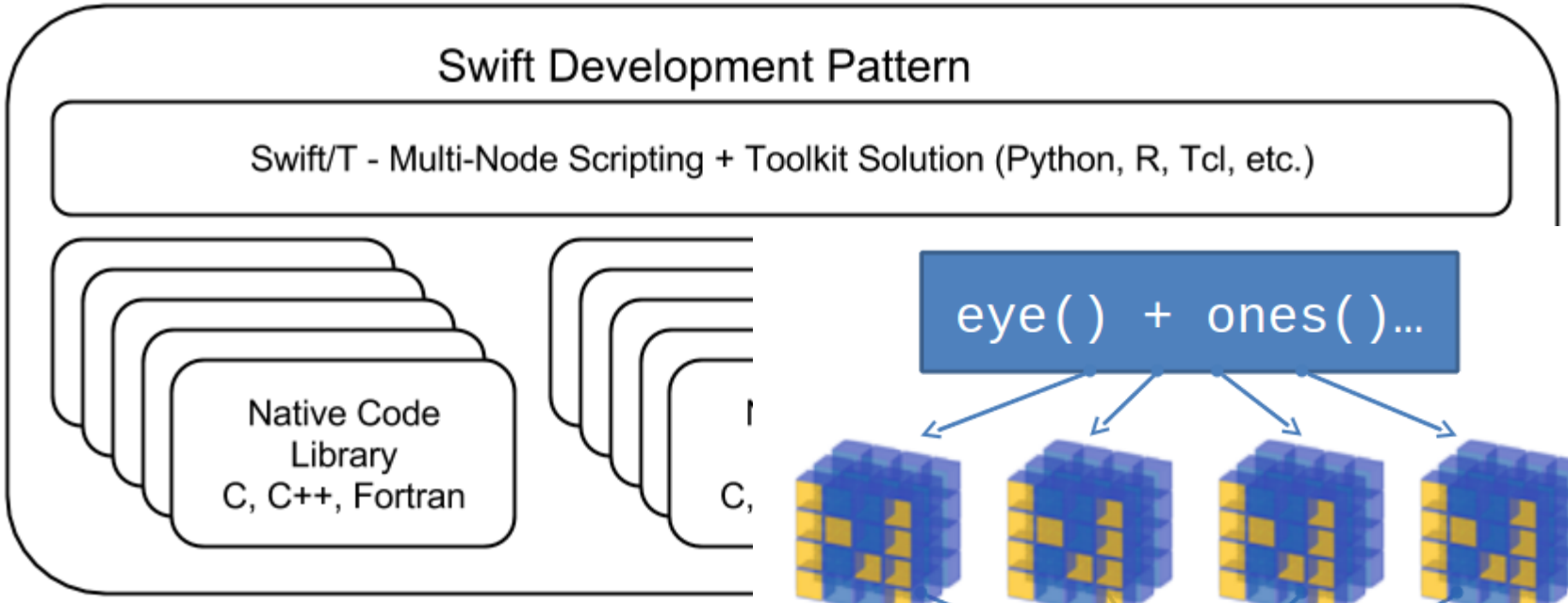
- Engines: evaluate dataflow operations

- Workers: execute tasks



- Wozniak et al. Turbine: A distributed-memory dataflow engine for high performance many-task applications. Fundamenta Informaticae 128(3), 2013

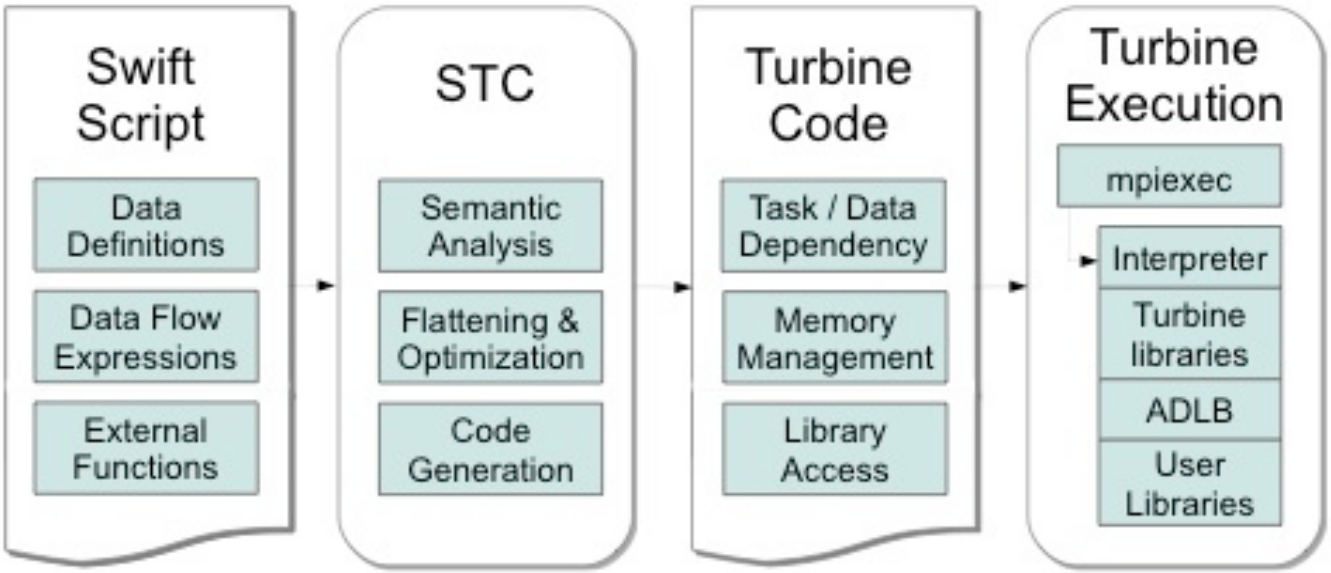
Support calls to embedded interpreters



We have plugins for Python, R, Tcl, Julia, and QtScript

- Wozniak et al. Toward computational experiment management via multi-language applications. Proc. ASCR SWP4XS, 2014.

STC: The Swift-Turbine Compiler



- STC translates high-level Swift expressions into low-level Turbine operations:
  - Create/Store/Retrieve typed data
  - Manage arrays
  - Manage data-dependent tasks

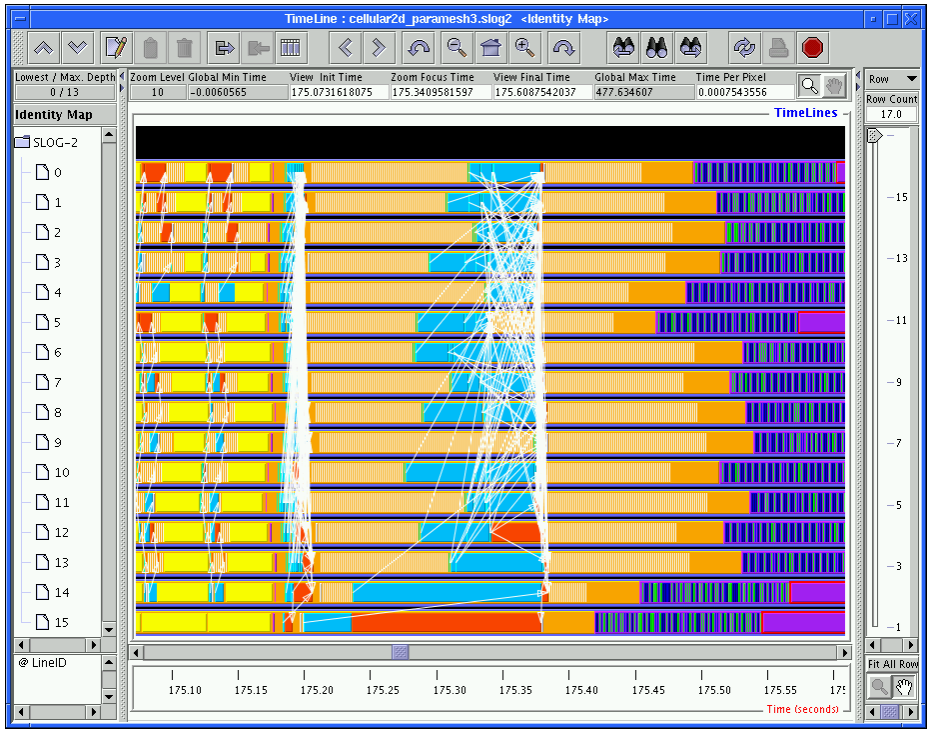
- Wozniak et al. Large-scale application composition via distributed-memory data flow processing. Proc. CCGrid 2013.
- Armstrong et al. Compiler techniques for massively scalable implicit task parallelism. Proc. SC 2014.

Logging and debugging in Swift

- Traditionally, Swift programs are debugged through the log or the TUI (text user interface)
- Logs were produced using normal methods, containing:
  - Variable names and values as set with respect to thread
  - Calls to Swift functions
  - Calls to application code
- A restart log could be produced to restart a large Swift run after certain fault conditions
- Methods require single Swift site: do not scale to larger runs

Logging in MPI

- The Message Passing Environment (MPE)
- Common approach to logging MPI programs
- Can log MPI calls or application events – can store arbitrary data
- Can visualize log with Jumpshot



- Partial logs are stored at the site of each process
  - Written as necessary to shared file system
    - in large blocks
    - in parallel
  - Results are merged into a big log file (CLOG, SLOG)

- Work has been done optimize the file format for various queries

Logging in Swift & MPI

- Now, combine it together
- Allows user to track down erroneous Swift program logic

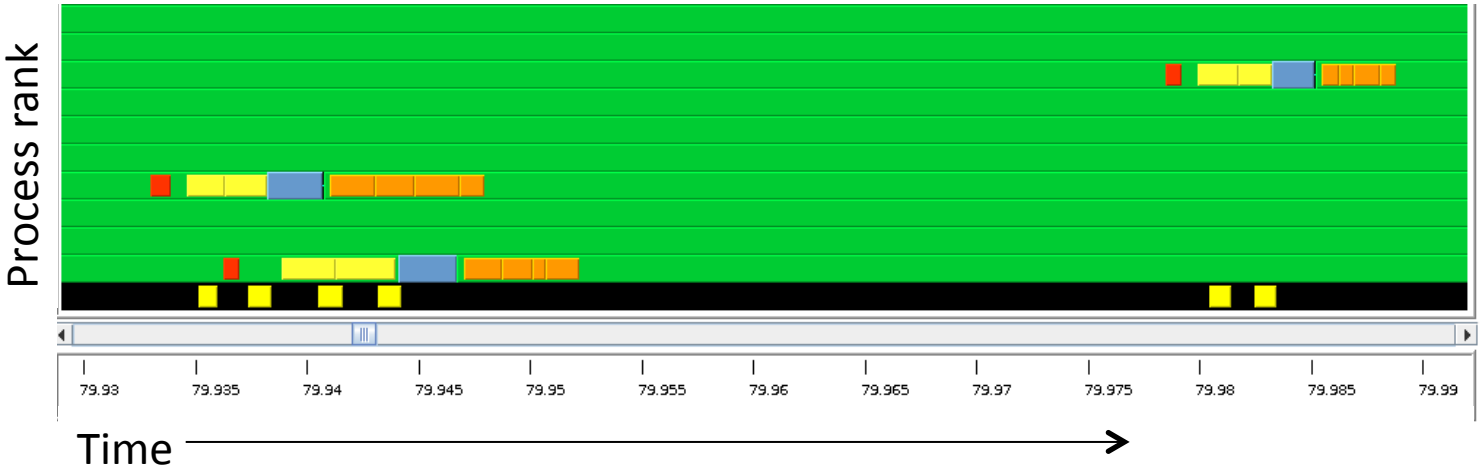
- Use MPE to log data, task operations, calls to native code
- Use MPE metadata to annotate events for later queries

- MPE **cannot** be used to debug native MPI programs that abort
  - On program abort, the MPE log is not flushed from the process-local cache
  - Cannot reconstruct final fatal events

- MPE **can** be used to debug Swift application programs that abort
  - We finalize MPE before aborting Swift
  - (Does not help much when developing Swift itself)
  - But primary use case is non-fatal arithmetic/logic errors

Visualization of Swift/T execution

- User writes and runs Swift script
- Notices that native application code is called with nonsensical inputs
- Turns on MPE logging – visualizes with MPE



Jumpshot view of PIPS application run

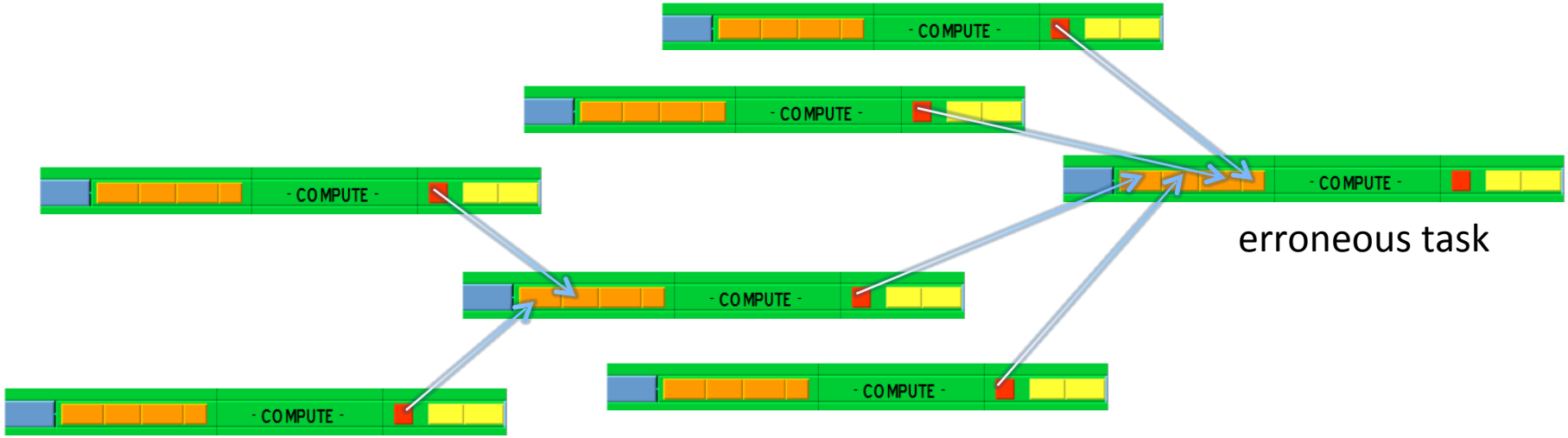
  - **PIPS task computation** **Store variable** **Notification (via control task)**
    Blue: Get next task **Retrieve variable**
    Server process (handling of control task is highlighted in yellow)

- Color cluster is task transition:

- Simpler than visualizing messaging pattern (which is not the user's code!)
- Represents Von Neumann computing model – load, compute, store

Debugging Swift/T execution

- Starting from GUI, user can identify erroneous task
  - Uses time and rank coordinates from task metadata
- Can identify variables used as task inputs
- Can trace provenance of those variables back in reverse dataflow



Aha! Found script defect. ← ← ← (searching backwards)

- Wozniak et al. A model for tracing and debugging large-scale task-parallel programs with MPE. Proc. LASH-C at PPoPP, 2013.

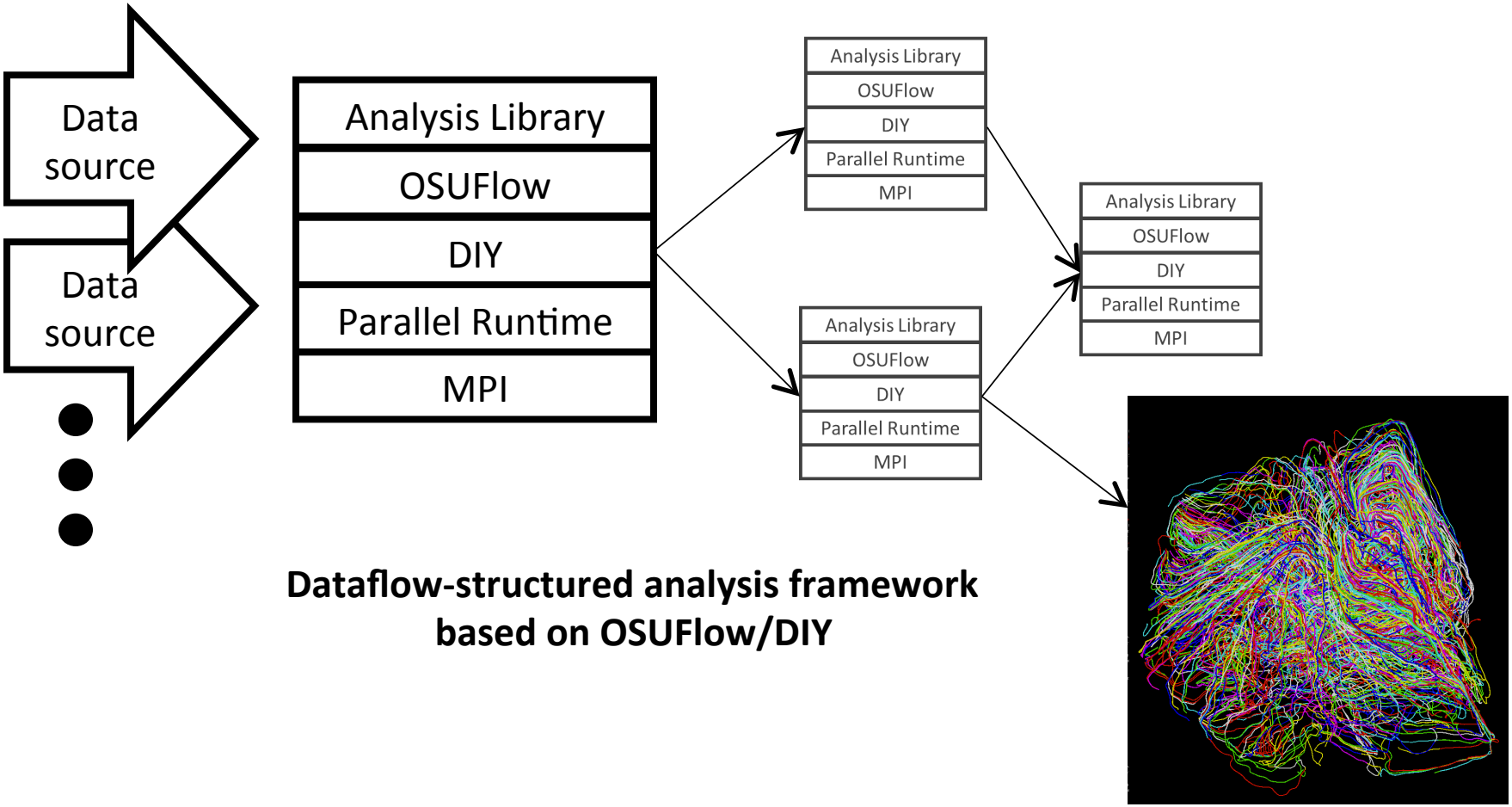
Other Swift/T features

- Task locality: Ability to send a task to a process
  - Allows for big data –type applications
  - Allows for stateful objects to remain resident in the workflow
  - `location L = find_data(D);`
    `int y = @location=L f(D, x);`
- Task priorities: Ability to set task priority
  - Useful for tweaking load balancing
- Updateable variables
  - Allow data to be modified after its initial write
  - Consumer tasks may receive original or updated values when they emerge from the work queue

- Wozniak et al. Language features for scalable distributed-memory dataflow computing. Proc. Dataflow Execution Models at PACT, 2014.

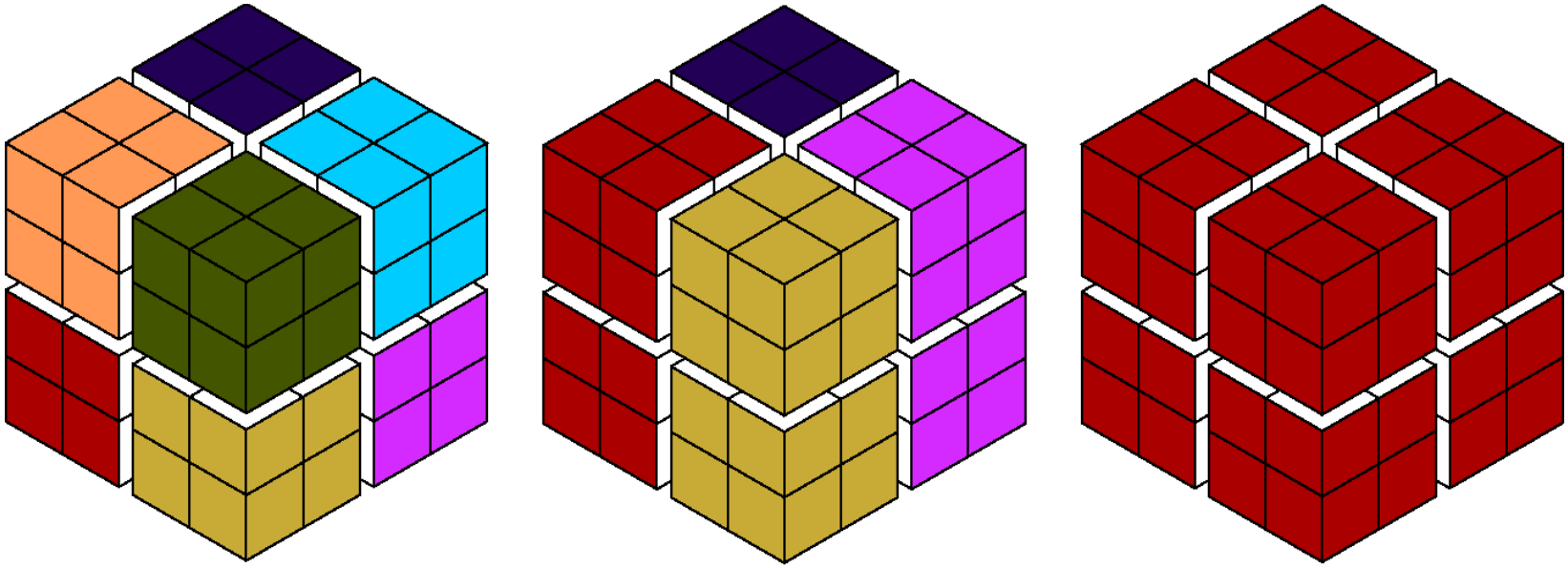
SWIFT/T: MPI TASKS

Dataflow+data-parallel analysis/visualization



Dataflow-structured analysis framework based on OSUFlow/DIY

Parameter optimization for data-parallel analysis: *Block factor*



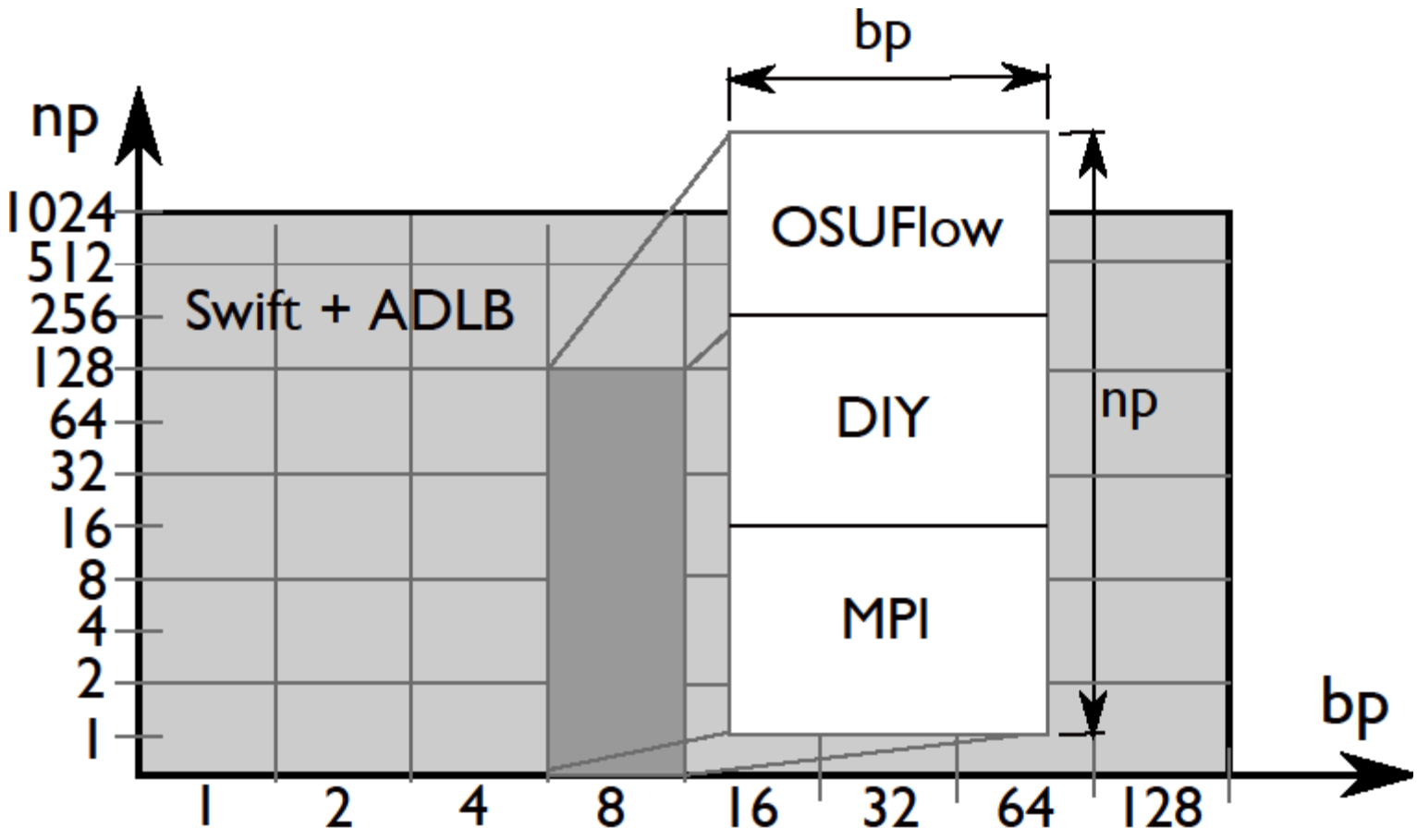
8 processes
1 block per process

4 processes
2 blocks per process

1 process
8 blocks per process

Can map blocks to processes in varying ways

Parameter optimization for data-parallel analysis: *Process configurations*



- **Try all configurations to find best performance**
- **Goal: Rapidly develop and execute sweep of MPI executions**

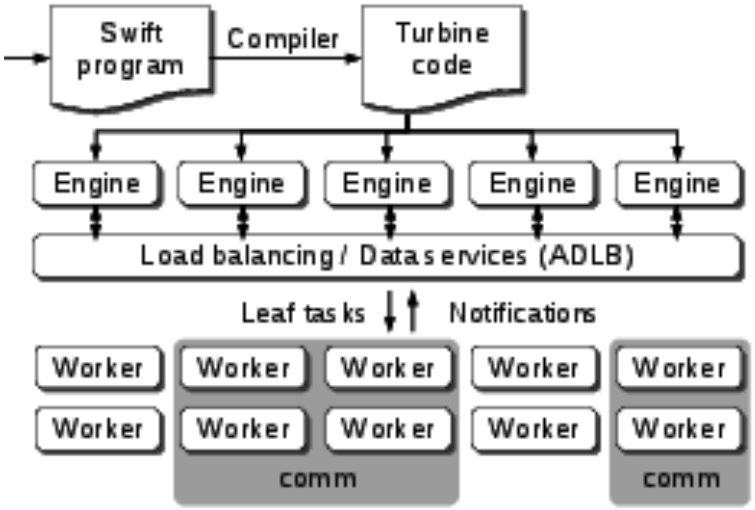
Refresher: MPI_Comm_create_group()

- In MPI 2, creating a subcommunicator was collective over the parent communicator
  - Required global coordination
  - Scalability concern
  - (Could use intercommunicator merges- somewhat slow)
- In MPI 3, the new `MPI_Comm_create_group()` allows the implementation to assemble the new communicator quickly from a group – only group members must participate
  - In ADLB, servers just pass rank list for new group to workers

- Motivating investigation by Dinan et al. identified fault tolerance and dynamic load balancing as key use cases – both relevant to Swift (Dinan et al., EuroMPI 2011.)

Parallel tasks in Swift/T

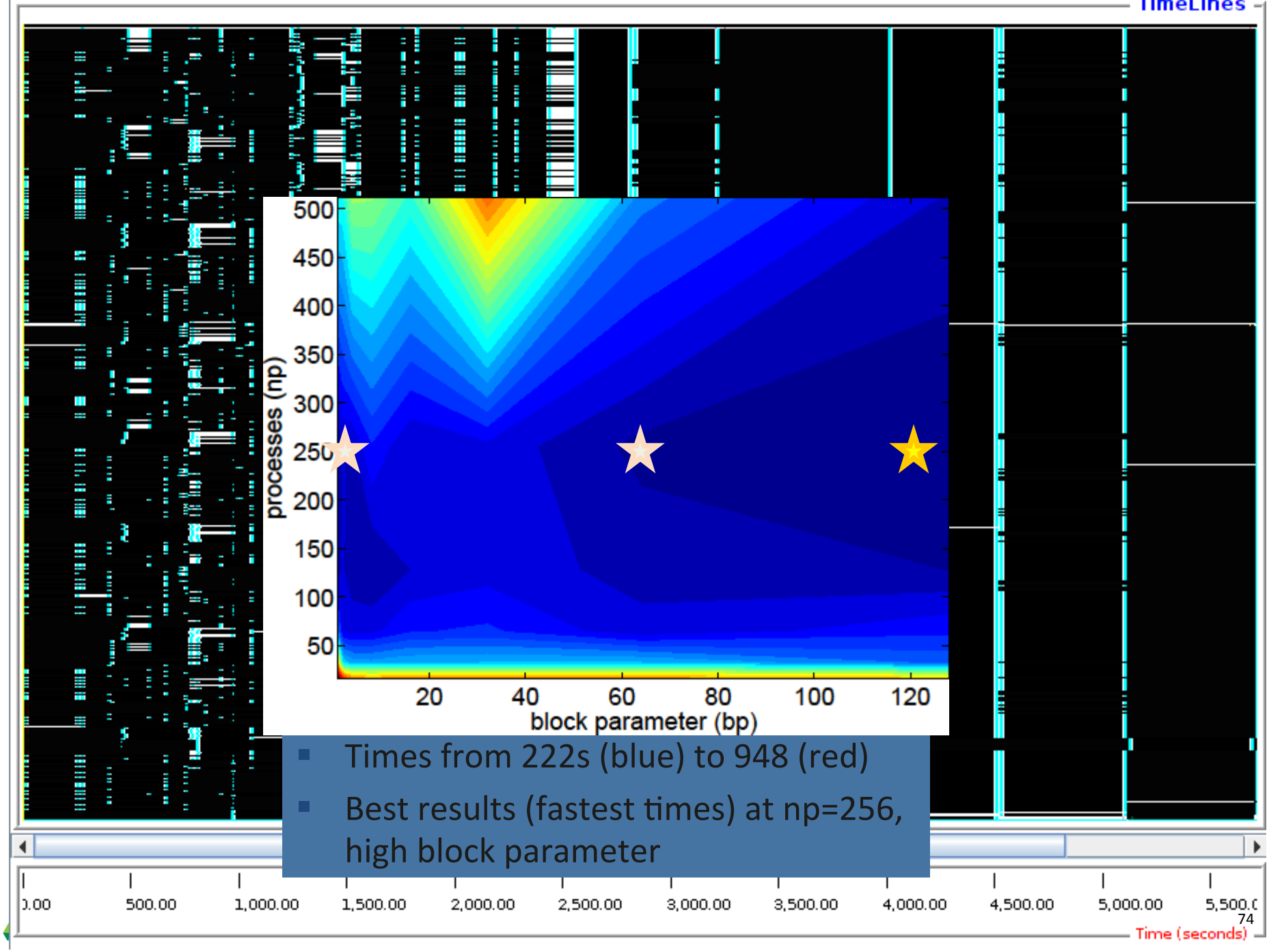
- Swift expression: `z = @par=8 f(x,y);`
- When x, y are stored, Turbine releases task `f` with `parallelism=8`
- Performs `ADLB_Put(f, parallelism=8)`
- Each worker performs `ADLB_Get(&task, &comm)`
- ADLB server finds 8 available workers
- Workers receive ranks from server
  - **Perform `MPI_Comm_create_group`**
- `ADLB_Get()` returns: `task=f, size(comm)=8`
- Workers perform user task
  - communicate on `comm`
- `comm` is released by Turbine



- Wozniak et al. Dataflow coordination of data-parallel tasks via MPI 3.0. Proc EuroMPI, 2013.

OSUFlow application

```
// Define call to OSUFlow feature MpiDraw
@par (float t) mpidraw(int bf) "mpidraw";

main {
  foreach b in [0:7] {
    // Block factor: 1-128
    bf = round(2**b);
    foreach n in [4:9] {
      // Number of processes/task: 16-512
      np = round(2**n);
      t = @par=np mpidraw(bf);
      printf("RESULT: bf=%i np=%i -> time=%0.3f",
                      bf,   np,      t);
    }}}
```



- Times from 222s (blue) to 948 (red)
- Best results (fastest times) at np=256, high block parameter

SWIFT/T APPLICATIONS

Swift integration into NAMD and VMD

www.ks.uiuc.edu/Research/swift

www.ks.uiuc.edu/Research/swift/

Apps | git-svn | RTM | BigBoard | Maps | Wikipedia | Swift, Inc | ProxyIt! | Other Bookmarks

NIH CENTER FOR MACROMOLECULAR MODELING & BIOINFORMATICS | UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Type Keywords SEARCH

THEORETICAL *and* COMPUTATIONAL BIOPHYSICS GROUP

Home | Research | Publications | Software | Instruction | News | Galleries | Facilities | About Us

- Home
- Overview
- Publications
- Research
  - Membrane Biology
  - Mechanobiology
  - Nanoengineering
  - Bioenergetics
  - SMD/IMD
  - Quantum Biology
  - Neurobiology
  - X-ray/Cryo-EM Modeling with MDFF
  - Driving Biomedical Projects
  - Collaborations
  - Other Topics
- Software
- Outreach

Integrating NAMD and VMD with Swift/T

NAMD and VMD have recently been successfully coupled to the **Swift/T high performance parallel scripting language** developed as part of the **ExM project**, a collaboration led by Argonne National Laboratory with University of Chicago and University of British Columbia, as a part of the **Department of Energy ASCR X-Stack** program. Swift/T is now supported as part of the **Swift project** under the **NSF SI2 program**. Standard NAMD 2.10 and VMD 1.9.2 binaries can be launched across the nodes of a parallel computer and efficiently execute Swift/T dataflow programs with functions implemented in the embedded Tcl scripting language. The NAMD and VMD user communities are already familiar with Tcl, and Tcl allows access to the two programs' complete functionality. The NAMD integration with Swift/T has been used to demonstrate n:m multiplexing of n replicas across a smaller arbitrary number m of NAMD processes, a very complex capability to implement with normal NAMD scripting that can be expressed naturally in under 100 lines of Swift/T code.

All example files: **directory**, **tar archive**

VMD Swift/T Hello World

VMD and Turbine must be built with compatible Tcl libraries so that VMD can dynamically load libtclturbine.so.

- Example command: `mpiexec -n 8 vmdwrapper -e vmdswift.tcl`
- Wrapper script to run standard VMD under MPI: **vmdwrapper**
- Tcl package and Swift startup for VMD: **vmdswift.tcl**
- Swift program source code: **hello.swift**
- Swift compiler Tcl output: **hello.tcl**

NAMD Swift/T Replica Exchange

NAMD and Turbine must be built with compatible Tcl libraries so that NAMD can dynamically load libtclturbine.so.

- Example command: `mpiexec -n 8 namdwrapper namdswift.tcl apoa1.namd --run 0 --source $cwd/replica.tcl < /dev/null &`
- Wrapper script to run multicore NAMD under MPI: **namdwrapper**
- Tcl package and Swift startup for NAMD: **namdswift.tcl**
- Swift program source code: **replica.swift**
- Swift compiler Tcl output: **replica.tcl**

NAMD Swift/T MPI Tight Binding

Charm++ and NAMD must be built from source code. An MPI-based Charm++ must be used. Apply the patches below to Charm++ and NAMD, respectively, to allow Turbine to access the Charm++ inter-partition communicator. Charm++, NAMD, and Turbine must be built with compatible Tcl and MPI libraries so that NAMD can dynamically load libtclturbine.so.

- Example command: `mpiexec -n 32 Linux-x86_64-g++.mpi/namd2 namdswift.tcl apoa1.namd --run 0 --source $cwd/replica.tcl +replicas 8 +stdout /var/tmp/stdout.%d.log < /dev/null &`
- Patch for Charm++ source code: **charmswift.patch**
- Patch for NAMD source code: **namdswift.patch**

Funded by a grant from the National Institute of General Medical Sciences of the National Institutes of Health

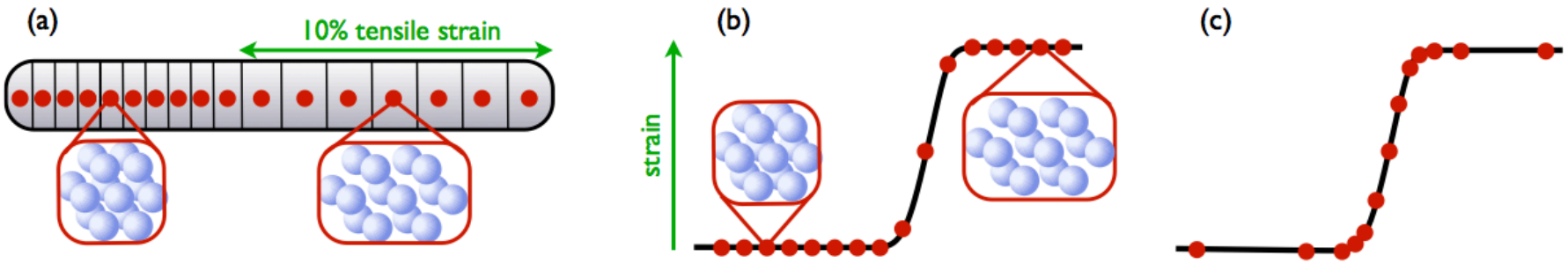
Center for the Physics of Living Cells

Beckman Institute for Advanced Science and Technology // National Institutes of Health // National Science Foundation // Physics, Computer Science, and Biophysics at University of Illinois at Urbana-Champaign

Contact Us // Material on this page is copyrighted; contact Webmaster for more information. // Document last modified on 10 Jul 2014 // 109 accesses since 25 Jun 2014 .

ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

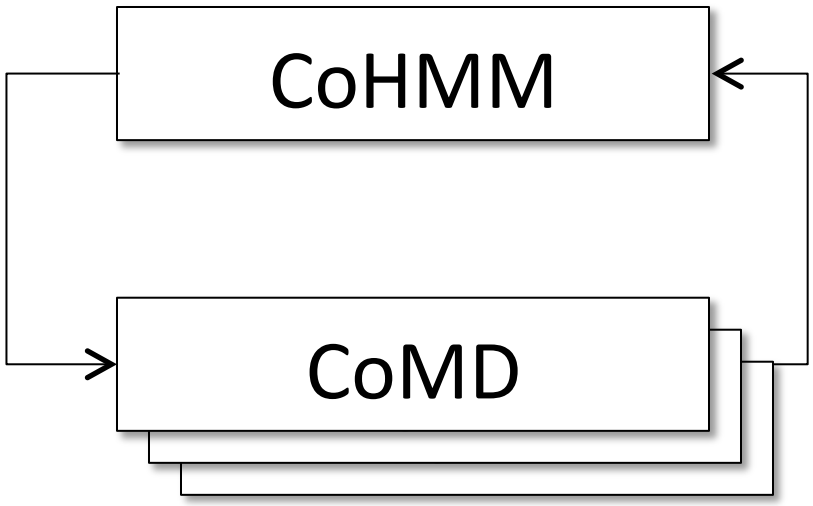
ExMatEx: Co-design for materials research



- CoHMM: Heterogeneous Multiscale Method
- CoMD: Molecular Dynamics
- Coarse-grain strain evolution using basic conservation laws
- Fine-grain molecular dynamics as necessary for physical coefficients

From http://www.exmatex.org

CoHMM/Swift



- Concurrency gained primarily by calls to CoMD

- 300 lines of sequential C
- Coordinates multiple sequential calls to CoMD
- We rewrote this in Swift

- 1000's lines of sequential C
- Simplified MD simulator
- Typically called as standalone program
- We exposed CoMD as a Swift function – no exec()

CoMD: Library access from Swift

- **CoMD binding: (example-1)**

```
string s = "-f data/8k.inp.gz";
int N = 3;
foreach i in [0:N-1] {
    float virial_stress = COMDSWIFT_runSim(s);
    printf("Swift: virial_stress: %e",
            virial_stress);
}
```

CoMD: Library access from CoHMM

C

```
#define ZERO_TEMP_COMD "../../CoMD/CoMD -x 6 -y 6 -z 6"
#ifdef ZERO_TEMP_COMD
// open pipe to CoMD
FILE *fPipe = popen(ZERO_TEMP_COMD,"r");
if (fPipe == NULL) {
        …
```

Swift

```
#define ZERO_TEMP_COMD "../../CoMD/CoMD -x 6 -y 6 -z 6"
#ifdef ZERO_TEMP_COMD
    string command = ZERO_TEMP_COMD;
    stressXX = COMDSWIFT_runSim(command);
#else
    // Just the derivative of the zero temp energy wrt A
    stressXX = rho0*c*c*(A-1);
#endif
```

CoHMM: Translation from C to Swift: main()

C

```
int main(int argc, char **argv) {
  initializedConservedFields();
  for (i = 0; i < 100; i++) {
    for (j = 0; j < 1; j++)
      fullStep();
```

Swift

```
main {
  (A[0], p[0], e[0]) = initializedConservedFields();
  for (int t = 0; t < 5; t = t+1) {
    (A[t+1], p[t+1], e[t+1]) =
                fullStep(A[t], p[t], e[t]);
```

CoHMM: Translation from C to Swift: call CoMD

C

```
void fluxes(double *A, double *p, double *e,
            double *f_A, double *f_p, double *f_e) {
    for (int i = 0; i < L; i++) {
        double stress = stressFn(A[i], e[i]);
        double v = p[i] / rho0;
        f_A[i] = -v;
        f_p[i] = -stress;
        f_e[i] = -stress*v;
```

Swift

```
(float f_A[], float f_p[], float f_e[])
fluxes(float A[], float p[], float e[]) {
  foreach i in [0:L-1] {
        float stress = stressFn(A[i], e[i]);
        float v = p[i] / rho0;
        f_A[i] = -v;
        f_p[i] = -stress;
        f_e[i] = -stress*v;
```

Can we build a Makefile in Swift?

- User wants to test a variety of compiler optimizations
- Compile set of codes under wide range of possible configurations
- Run each compiled code to obtain performance numbers
- Run this at large scale on a supercomputer (Cray XE6)

- **In Make you say:**

```
CFLAGS = ...
f.o : f.c
    gcc $(CFLAGS) f.c -o f.o
```

In Swift you say:

```
string cflags[] = ...;
f_o = gcc(f_c, cflags);
```

CHEW example code

Apps

```
app (object_file o) gcc(c_file c, string cflags[]) {
// Example:
//  gcc   -c   -O2    -o  f.o f.c
   "gcc" "-c" cflags "-o" o   c;
}

app (x_file x) ld(object_file o[], string ldflags[]) {
// Example:
//  gcc          -o  f.x f1.o f2.o ...
   "gcc" ldflags "-o" x   o;
}

app (output_file o) run(x_file x) {
  "sh" "-c" x @stdout=o;
}

app (timing_file t) extract(output_file o) {
  "tail" "-1" o "|" "cut" "-f" "2" "-d" " " @stdout=t;
}
```

Swift code

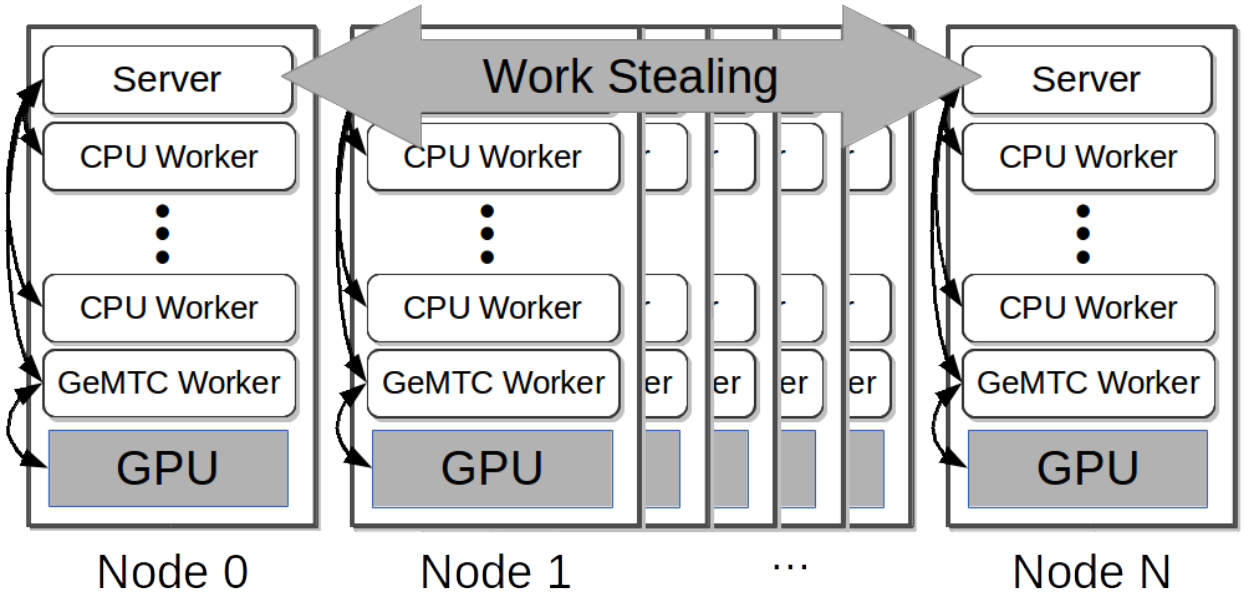
```
string program_name = "programs/program1.c";
c_file c = input(program_name);

// For each
foreach O_level in [0:3]  {
  make file names…
  // Construct compiler flags
  string O_flag = sprintf("-O%i", O_level);
  string cflags[] = [ "-fPIC", O_flag ];

  object_file o<my_object> = gcc(c, cflags);
  object_file objects[] = [ o ];
  string ldflags[] = [];
  // Link the program
  x_file x<my_executable> = ld(objects, ldflags);
  // Run the program
  output_file out<my_output> = run(x);
  // Extract the run time from the program output
  timing_file t<my_time> = extract(out);
```

Swift Use of GPUs

GeMTC: GPU-enabled Many-Task Computing

Approach:

1) Deploy kernel to manage GPU warps
2) Manage memory
3) Integrate with workflow system (Swift/T)



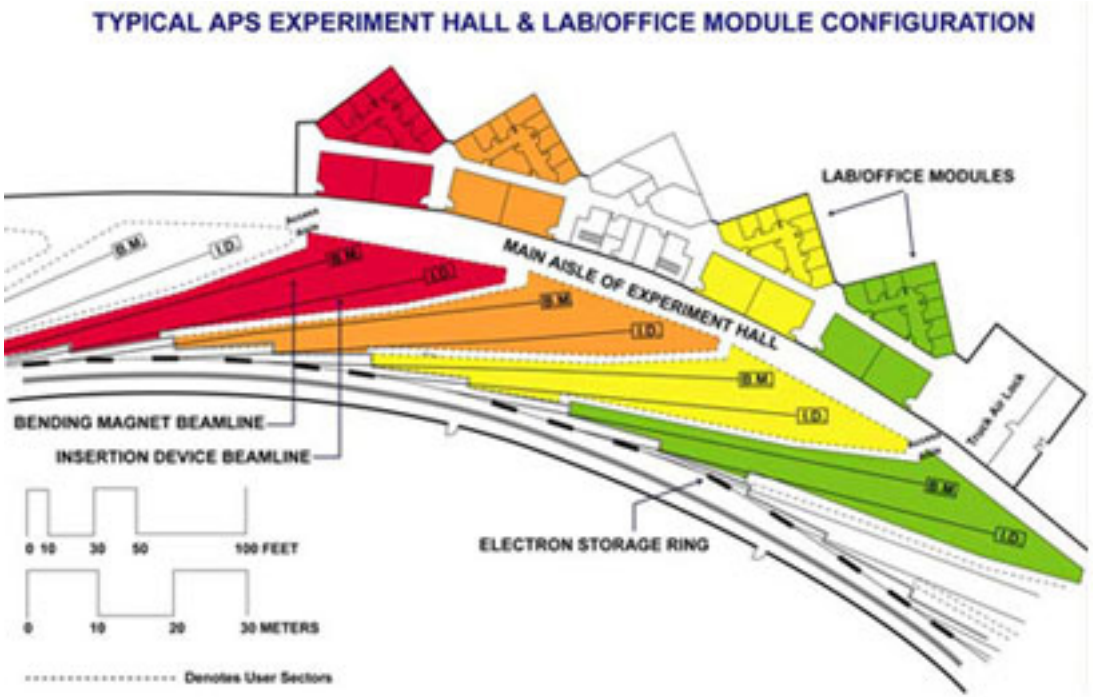
- Krieder et al. Evaluation of Many-Task Computing on Accelerators for High-End Systems. Proc. HPDC 2014.

DISCOVERY ENGINES LDRD: WORKFLOWS

Advanced Photon Source (APS)

Advanced Photon Source (APS)

- Moves electrons at electrons at >99.999999% of the speed of light.
- Magnets bend electron trajectories, producing x-rays, highly focused onto a small area
- X-rays strike targets in 35 different laboratories – each a lead-lined, radiation-proof experiment station



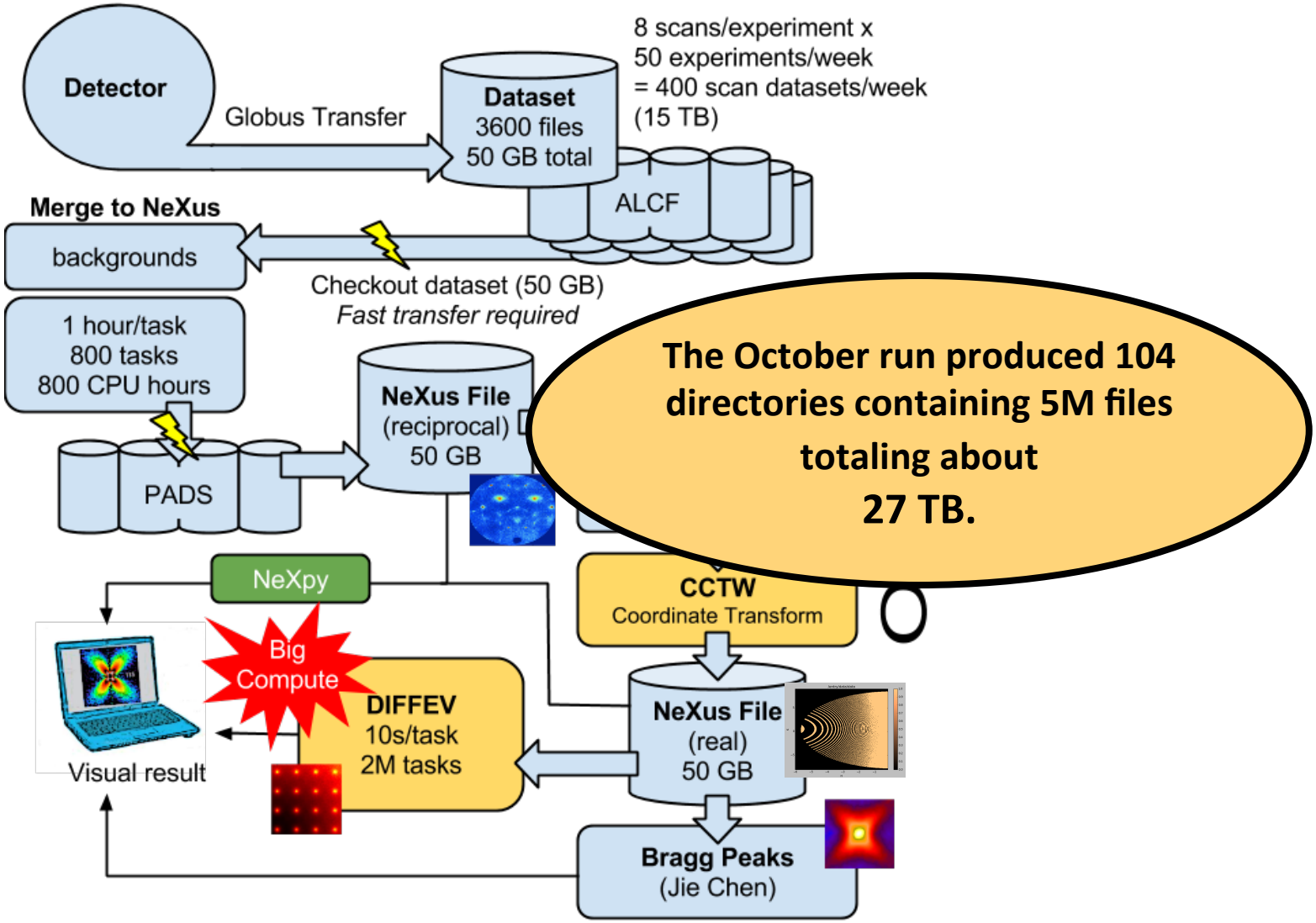
Data management for the energy sciences

- "Despite the central role of **digital data** in Dept. of Energy (DOE) research, the methods used to manage these data and to support the information and **collaboration processes** that underpin DOE research are often **surprisingly primitive**…"

  - DOE Workshop Report on Scientific Collaborations (2011)

- Our goals:
  - Modify the operating systems of APS stations to allow real-time streaming to a novel data storage/analysis platform.
  - Converting data from the standard detector formats (usually TIFF) to HDF5 and adding metadata and provenance, based on the NeXus data format.
  - Rewrite analysis operations to work in a massively parallel environment.
  - Scale up simulation codes that complement analysis.

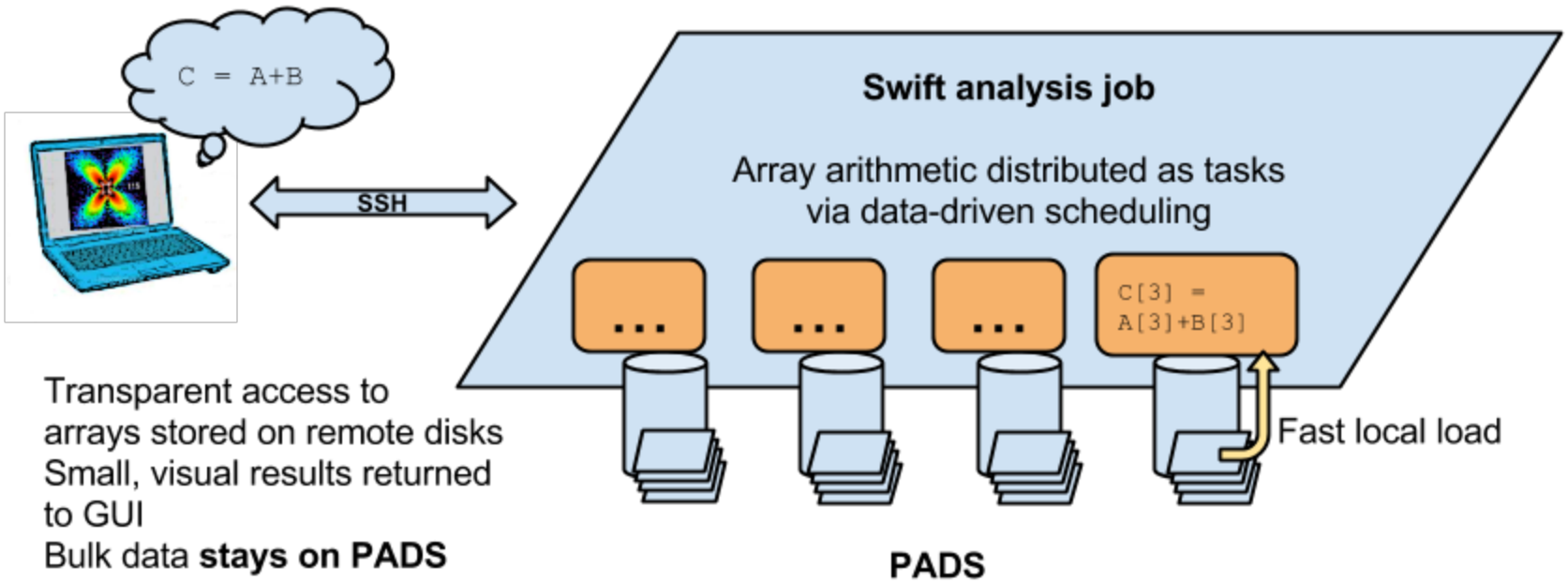
Data ingest/analysis/archive



PADS: Petascale Active Data Store

- 23 higher-end nodes for data-intensive computing, repurposed for this work (installed in 2009)
  - Each node has 12-way RAID for very fast local disk operations
- Previously, difficult to use as "Active Data Store"
  - Difficult to access specific nodes through PBS scheduler
  - No catalog (where is my data?)
  - *No way to organize/access Data Store!*
- Solution: Swift/T
  - Organizes distributed data using Swift data structures and mappers
  - Leaves data on nodes for later access
  - Allows for targeted tasks (can send work to node with data chunk)
  - Integrates with Globus Catalog for metadata, provenance, archive...
  - Combining unscheduled resource access with high performance data rates will allow for **real-time beamline data analysis, accelerating progress for materials science efforts**

Interactive analysis powered by scalable storage

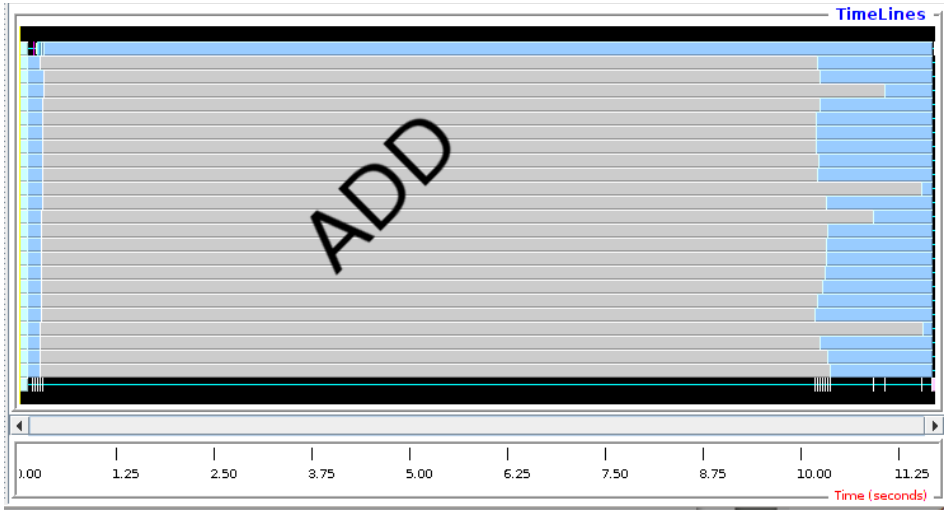


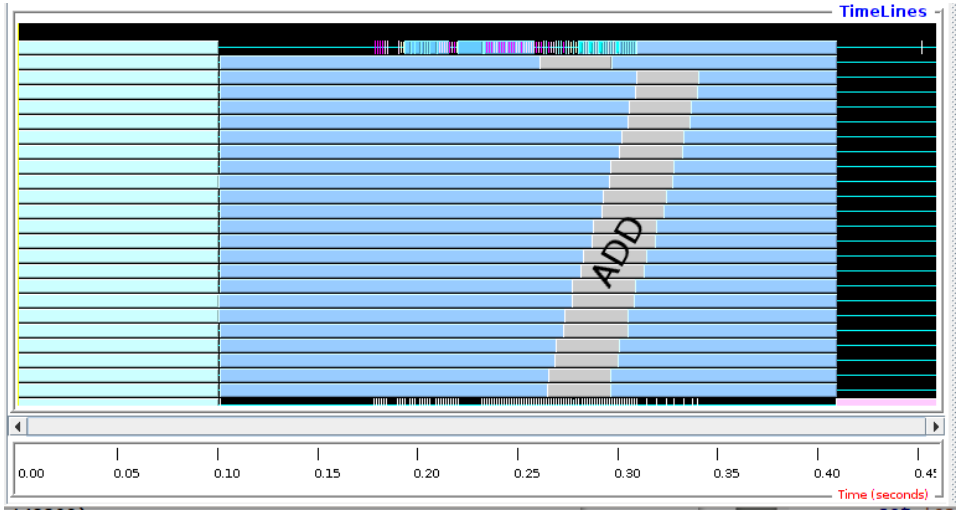
- Transparent access to arrays stored on remote disks
- Small, visual results returned to GUI
- Bulk data **stays on PADS**

- Replace GUI analysis internals with operations on remote data

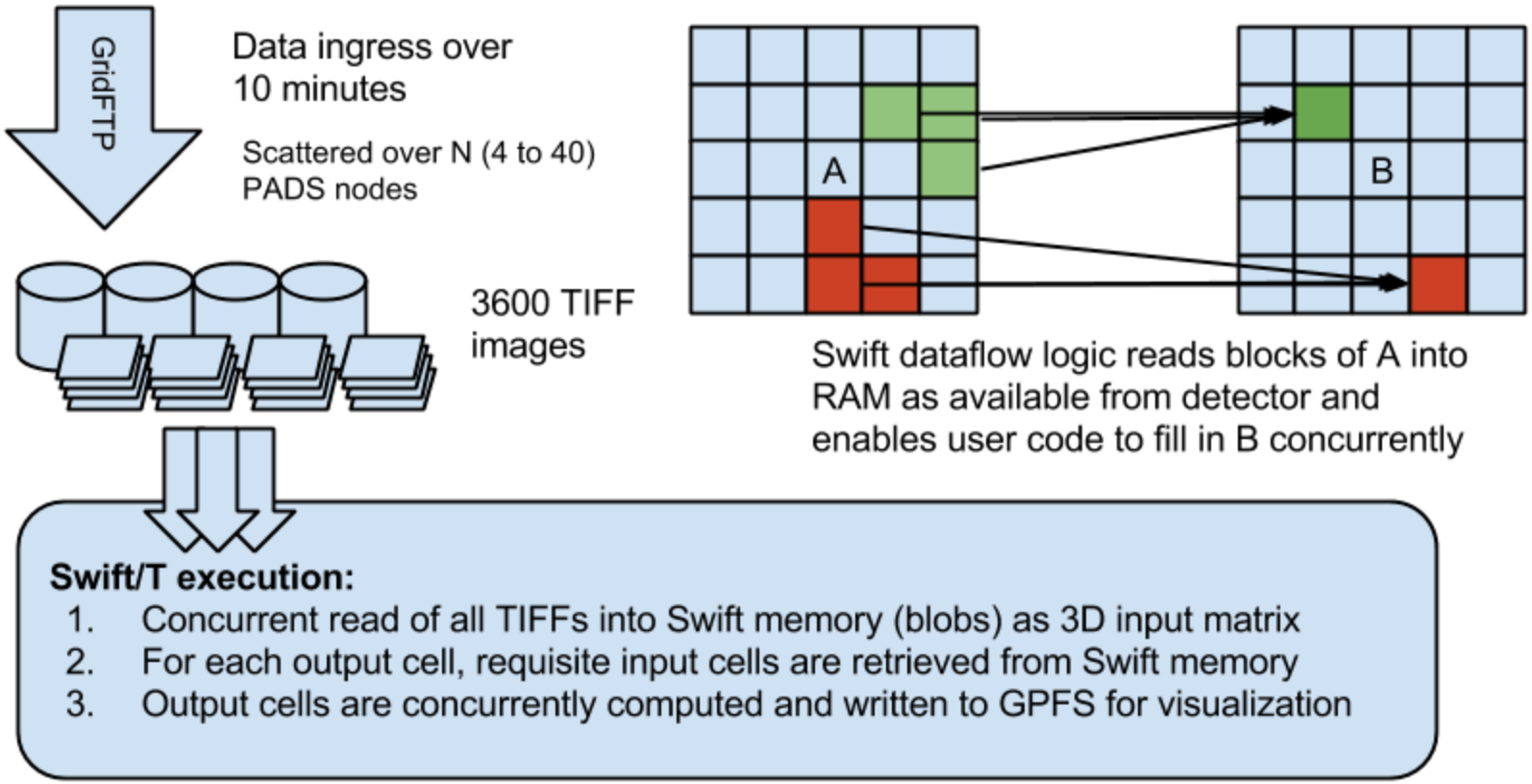
Remote matrix arithmetic: Initial results

- Initial run shows performance issue: addition took too long
- Swift profiling isolated issue: convert addition routine from script to C function: obtained 10,000 X speedup
- Swift/T integrates with MPE/ Jumpshot and other MPI-based performance analysis techniques





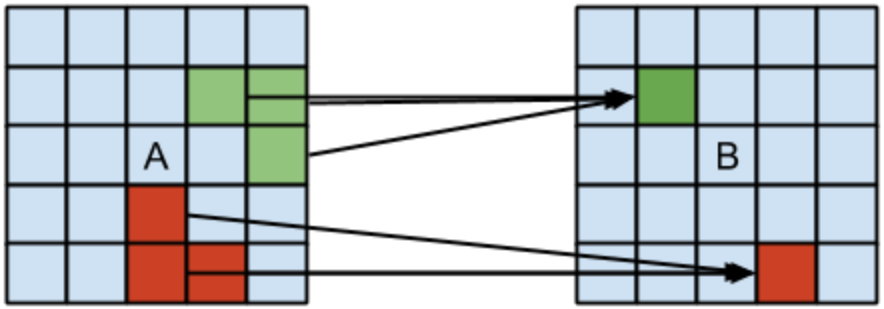
Crystal Coordinate Transformation Workflow



MapReduce-like pattern expressed elegantly in Swift

CCTW: Swift/T application (C++)

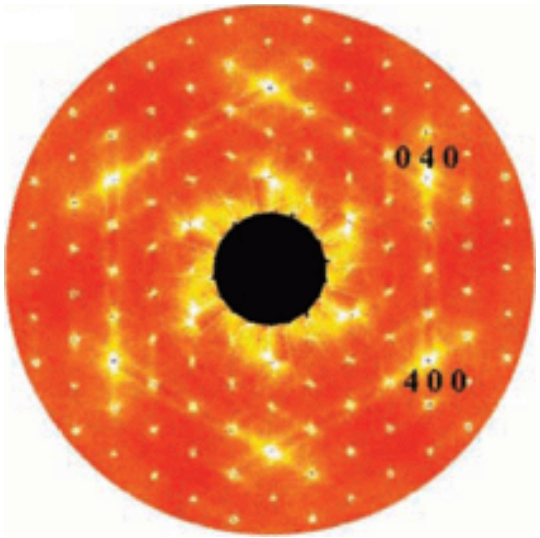
```
bag<blob> M[];
foreach i in [1:n] {
    blob b1= cctw_input("pznpt.nxs");
    blob b2[];
    int outputId[];
    (outputId, b2) = cctw_transform(i, b1);
    foreach b, j in b2 {
        int slot = outputId[j];
        M[slot] += b;
    }}
    foreach g in M {
        blob b = cctw_merge(g);
        cctw_write(b);
    }}
```



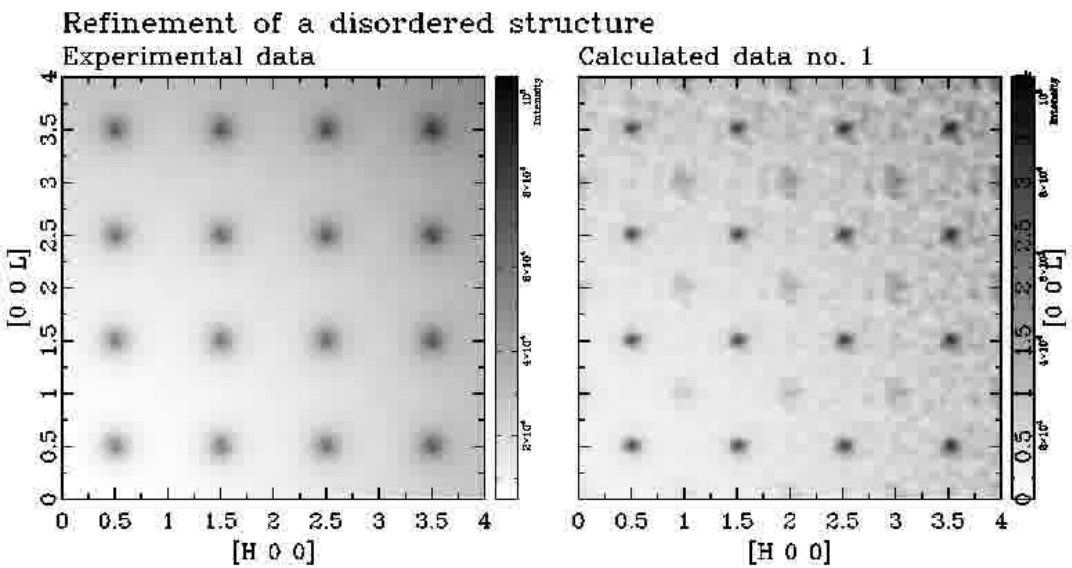
Diffuse scattering and crystal analysis

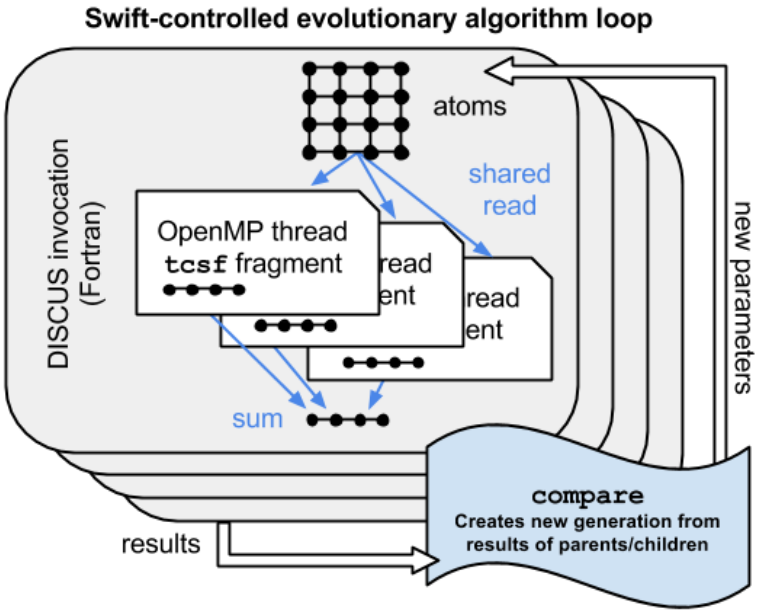
- DISCUS is a general program to generate disordered atomic structures and compute the corresponding experimental data such as single crystal diffuse scattering (http://discus.sourceforge.net)
- Given experimental data, can we fit a modeled crystal to the measurement?

- Experimental image: (Billinge, 2006)



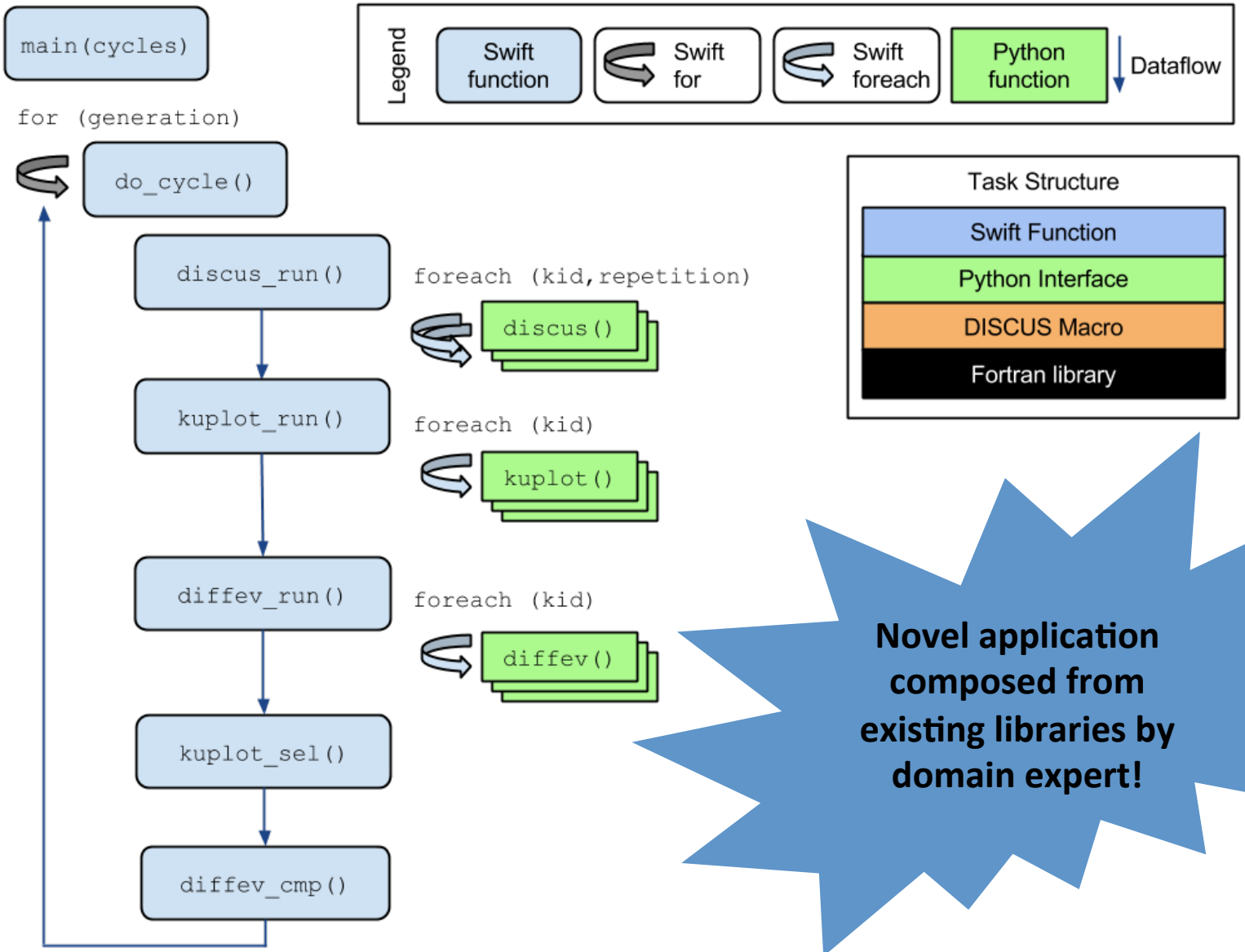
DIFFEV: Scaling crystal diffraction simulation





- Determines crystal configuration that produced given scattering image through simulation and evolutionary algorithm
- Swift/T calls DISCUS via Python interfaces

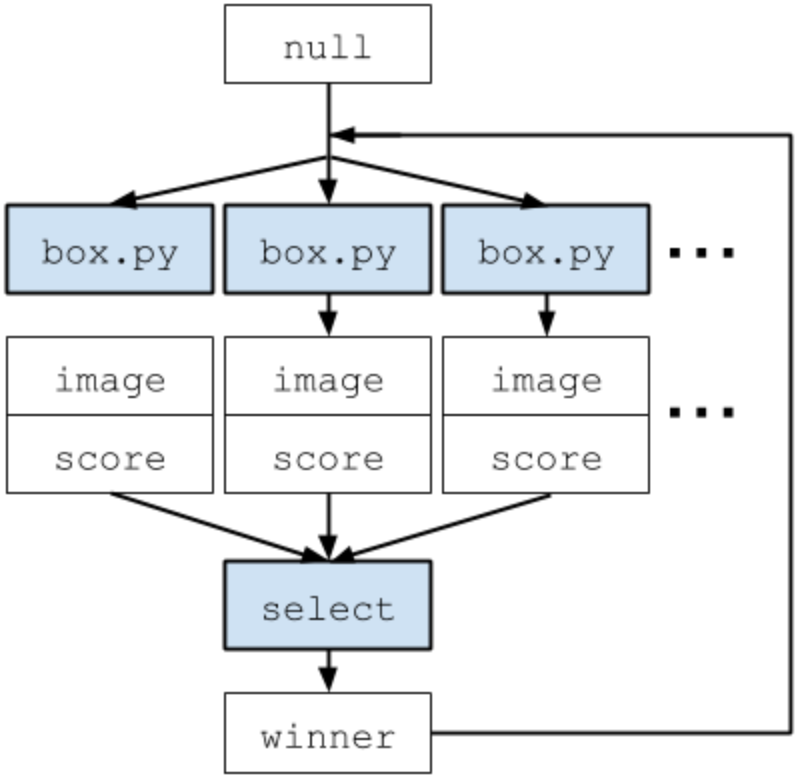
DIFFEV: Genetic algorithm via dataflow



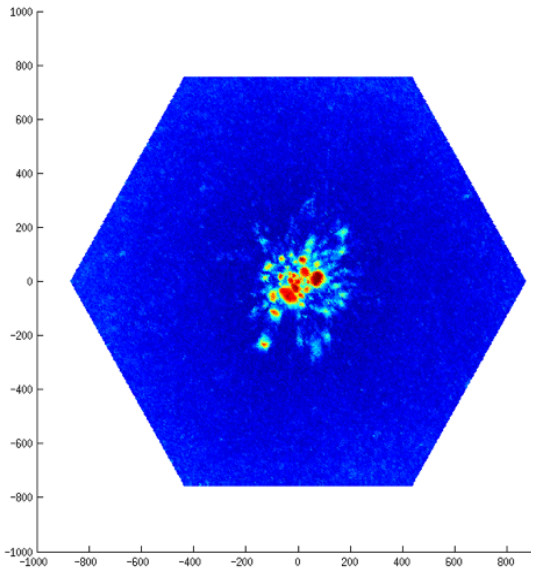
Novel application composed from existing libraries by domain expert!

R. Harder workflow: Genetic algorithm

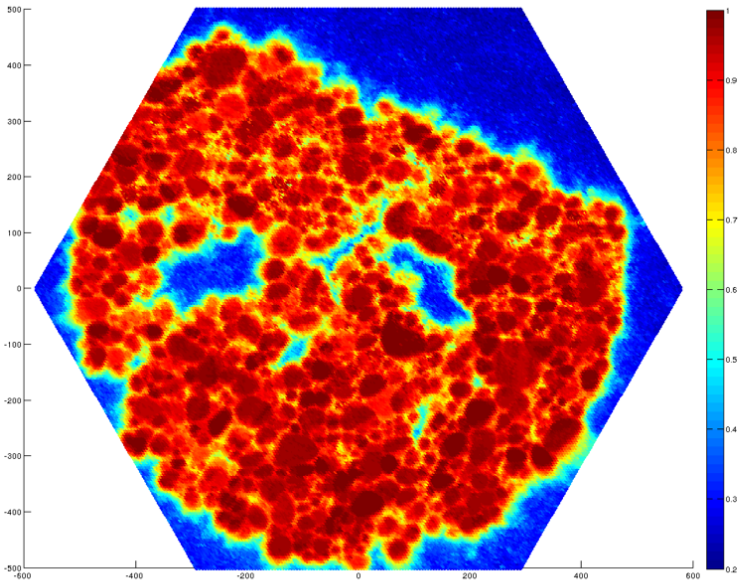
```
individuals = toint(argv("individuals"));
ngenerations = toint(argv("ngenerations"));
file winners[];
winners[0] = input("null.winner");
for (int generation = 1; generation < ngenerations;
                                generation = generation+1) {
  file population[];
  foreach box_index in [0:individuals-1] {
    file d<sprintf("d-%i-%i.out",generation,box_index)>;
    file s<sprintf("d-%i-%i.score",generation,box_index)>;
    (d,s) = box(box_index, generation, winners[generation-1]);
    population[box_index] = d;
  }
  file winner_file<sprintf("d-%i.winner", generation)> =
                 select(generation, population);
  winners[generation] = winner_file;
 }}
```



High-Energy Diffraction Microscopy



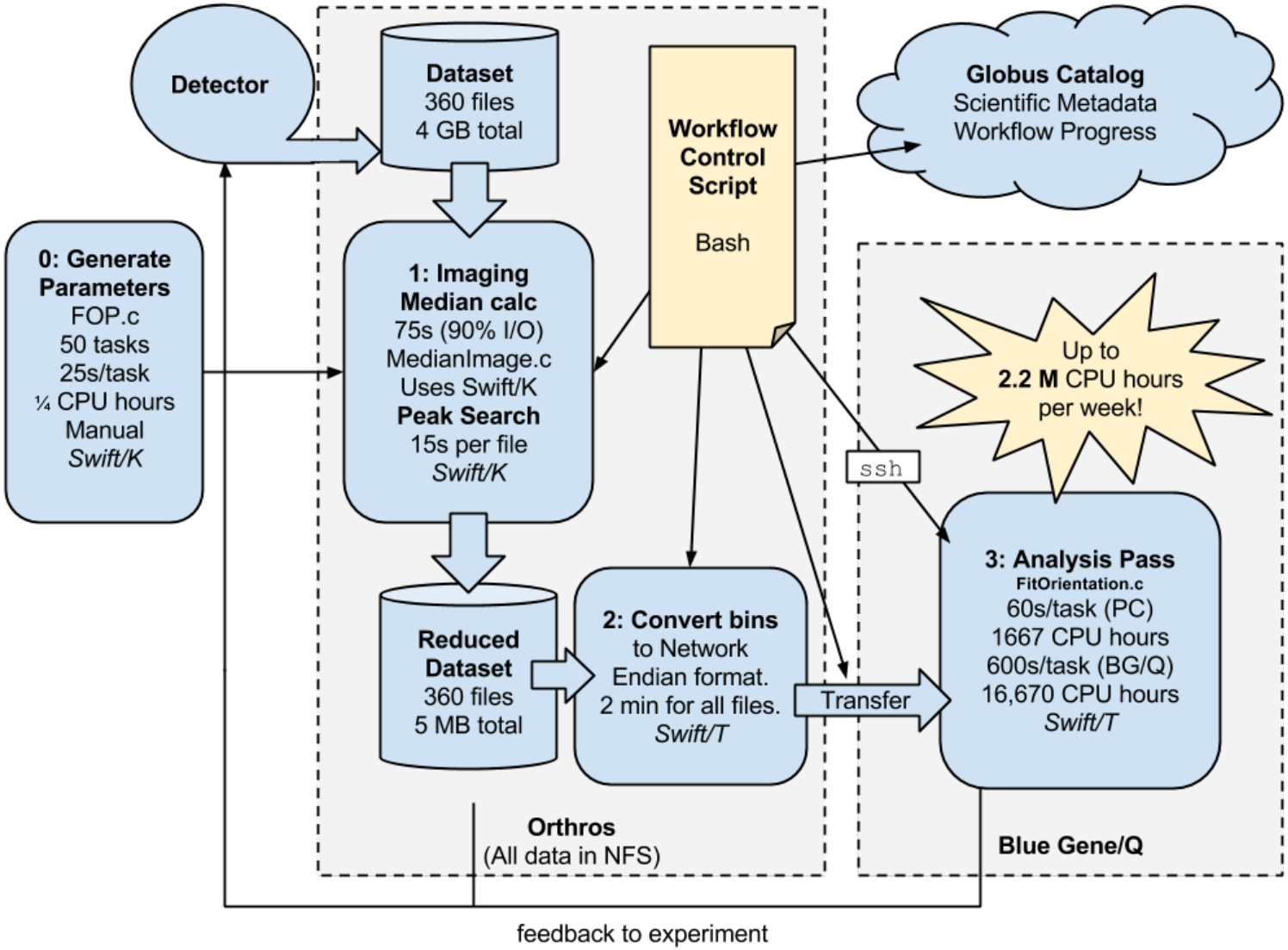
October 2013: Without Swift



April 2014: With Swift

- Near-field high-energy diffraction microscopy discovers metal grain shapes and structures
- The experimental results are greatly improved with the application of Swift-based cluster computing (**RED** indicates higher confidence in results)

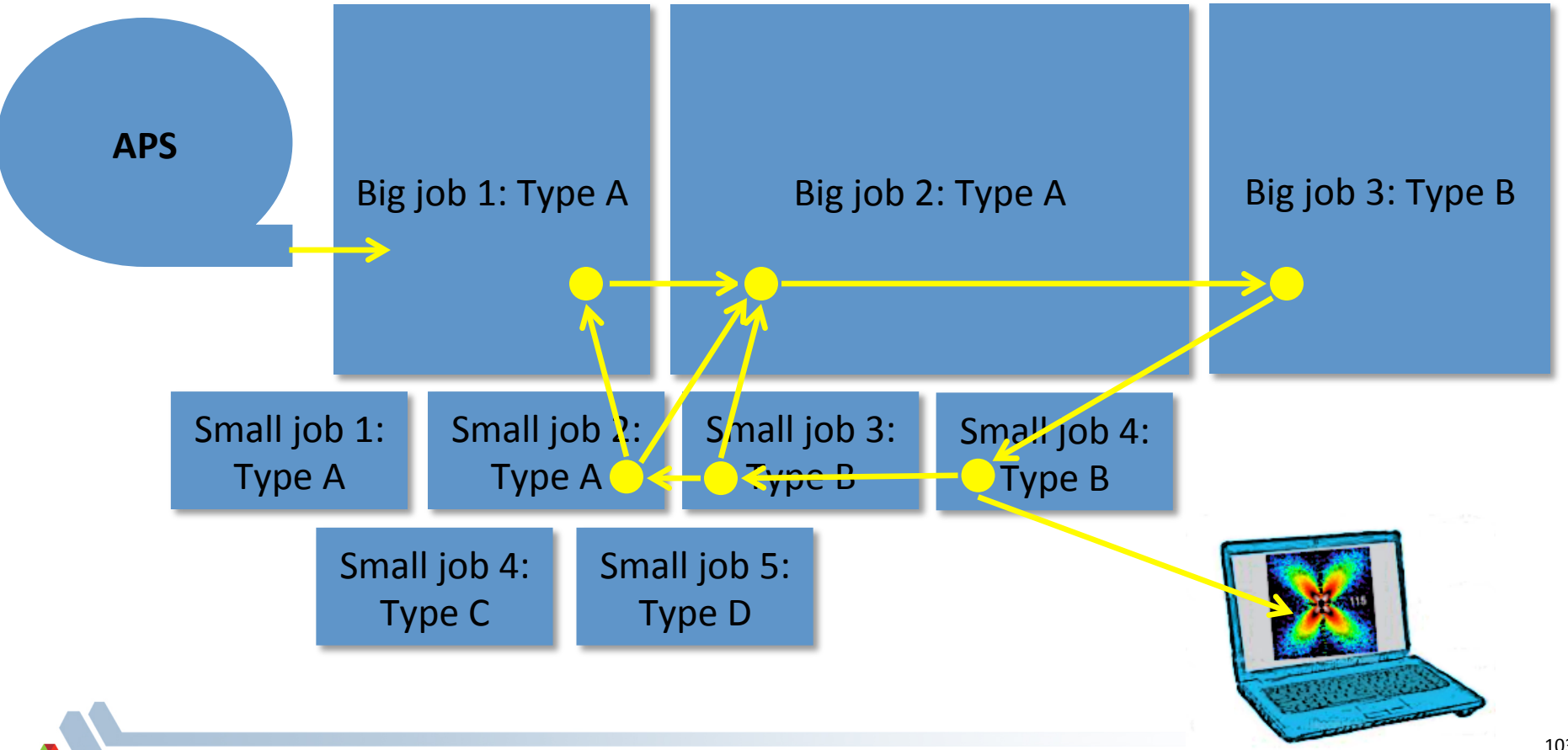
NF-HEDM: Cross-lab workflow



FUTURE WORK

Extreme scale application ensembles

- Develop Swift for exascale experiment ensembles
  - Deploy stateful, varying sized jobs
  - Outermost, experiment-level coordination via dataflow
  - Plug in experiments and human-in-the-loop models (dataflow filters)



Future Work

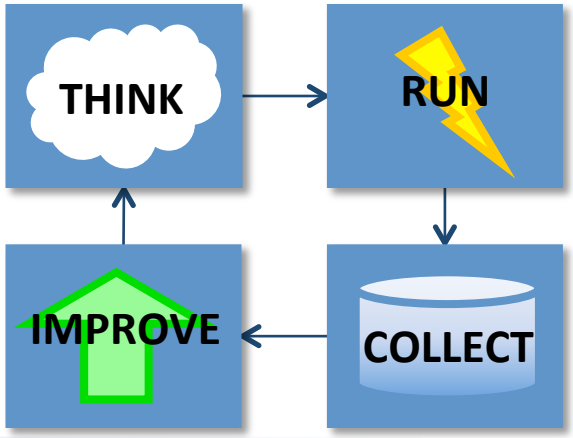
- Develop Swift for exascale
  - Continue scaling work: Study distributed dataflow for realistic patterns
  - Ease integration with native code

- Application collaborations
  - Materials science: APS (Osborn, Sharma)
  - Molecular dynamics: NAMD (Phillips), LAMMPS (Whitmer)

- Connect with novel systems elsewhere in MCS, ALCF:
  - Memcached (Isaila et al.)
  - Tess (Peterka et al.)
  - Filesystems (Ross et al.)

- Connect with new applications at the CI and elsewhere!

Summary

- Swift: High-level scripting for outermost programming constructs
  - Handles many aspects of the scientific computing experience
  - Described how dataflow logic is distributed
  - New features for parallel tasks
- Thanks to the Swift team: Mike Wilde, Ketan Maheshwari, Tim Armstrong, David Kelly, Yadu Nand, Mihael Hategan, Scott Krieder, Ioan Raicu, Dan Katz, Ian Foster
- Thanks to project collaborators: Tom Peterka, Jim Dinan, Ray Osborn, Reinhard Neder, Guy Jennings, Hemant Sharma, Rachana Ananthakrishnan, Ben Blaiszik, Kyle Chard, Tim Germann, and others

- **Questions?**