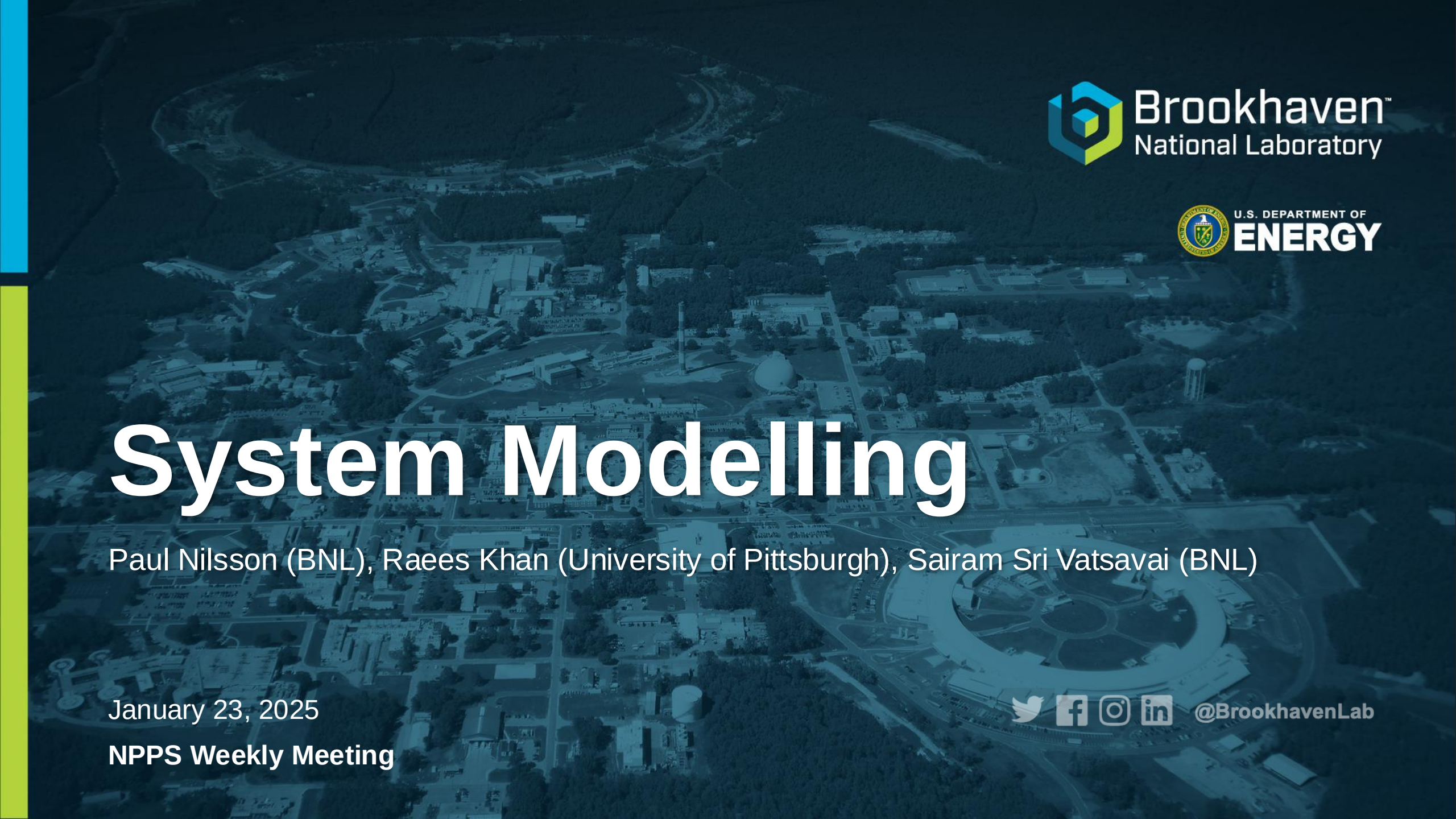


Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

System Modelling

Paul Nilsson (BNL), Raees Khan (University of Pittsburgh), Sairam Sri Vatsavai (BNL)

January 23, 2025

NPPS Weekly Meeting

@BrookhavenLab

Introduction (and reminders)

- REDWOOD is an ongoing project that started in 2023
- A main objective is the optimization of Workflow Management System (WFMS) to enhance system resilience
- Two targets for optimization
  - Brokerage – where should the job go?
  - Data placement – where should the data go?
- Different approaches will be used, including ML techniques
- This talk will discuss how WFMS could be modelled and simulated
  - Different algorithms for optimization could then be studied

Simulators

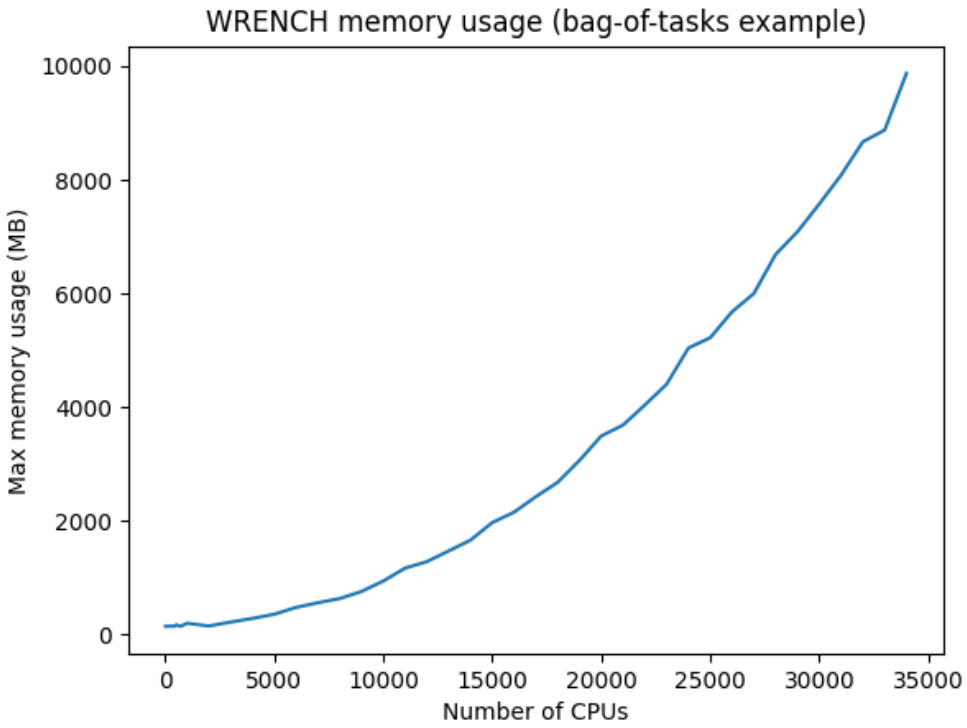
Which tools are or interest?

A well-known framework for developing simulators of distributed applications targeting distributed platforms, which can in turn be used to prototype, evaluate and compare relevant platform configurations, system designs, and algorithmic approaches. In active development for over 20 years

An open-source framework designed to make it easy for users to develop accurate and scalable simulators of distributed computing applications, systems, and platforms. WRENCH is built op top of SimGrid.

WRENCH

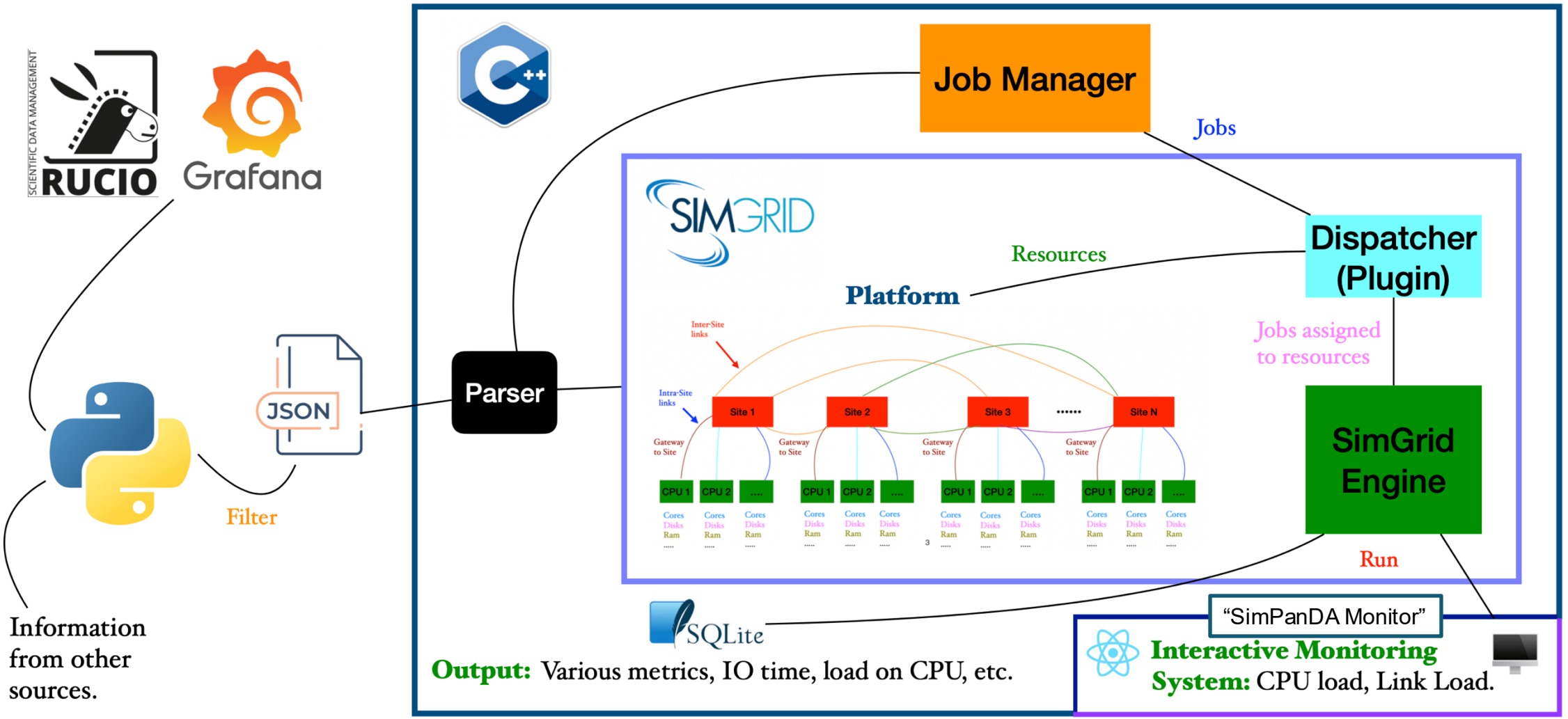
- WRENCH would be a good choice for building a simulator as it was developed to reduce the complexity of SimGrid, and make things simpler for the developer
- However, testing showed that it is too memory hungry for a larger amount of CPUs
  - Modified standard example
    - Dynamically creating simple geometry for single computing resource
  - Dramatic increase of used memory as number of hosts increase
  - Site of BNL size needs around 8 GB to run simulation
    - I.e. 150+ grid sites won't be possible
  - Most of the memory is used by geometry, another 10-20% increase in memory usage when simulation starts (10,000 jobs)
- Other projects also had "difficulties" with WRENCH (DCSim)
  - Could still be of interest in case of simpler studies, since e.g. it comes with a simulation of Condor



SimGrid

- Simulator for ATLAS using latest version of SimGrid is in development since half a year
  - Code repository https://github.com/REDWOOD24/ATLAS-GRID-SIMULATION
- Topology generated from JSON files using metrics data from Rucio and from other sources
  - Connections between sites and bandwidths used to construct grid of sites
  - Number of CPUs/cores extracted from job records using data reported by the pilot
  - Scaled corepower (hs06) used as GFLOPS (not proportional but corepower is the best we have)
- Trivial job generator used so far
  - Jobs have so far been generated from a gaussian distribution
  - Also, jobs can be generated using historical data
- A study of memory usage as function of number of CPUs looks good
  - Memory usage did not pass 3 GB while running 5k tasks with full ATLAS grid geometry
- Error handling to be implemented
  - I.e. currently 100% of jobs succeeds

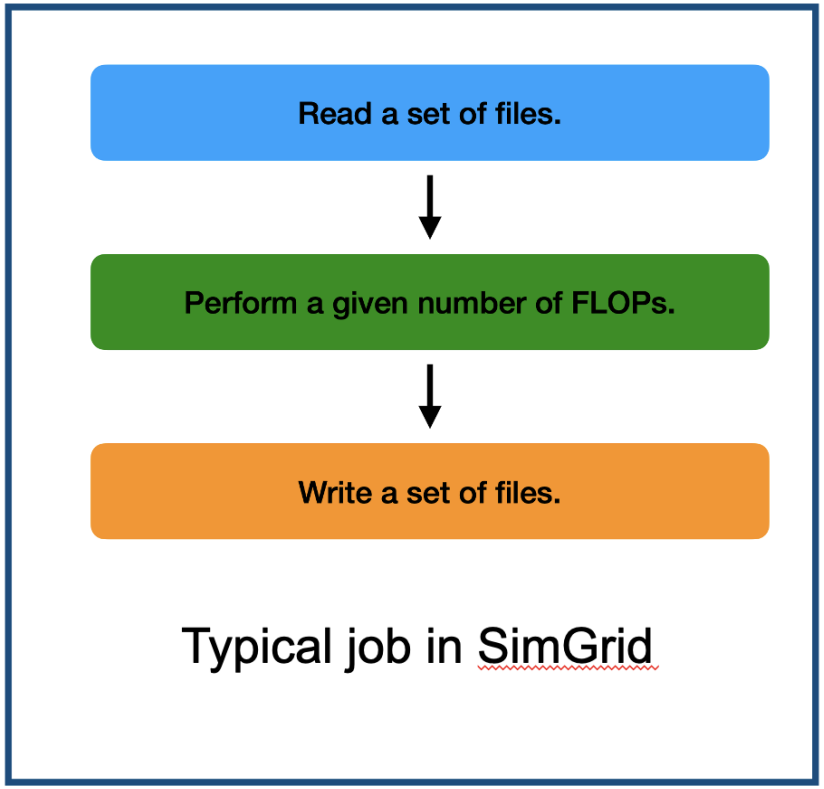
Simulator Workflow



R. Khan

Dispatcher

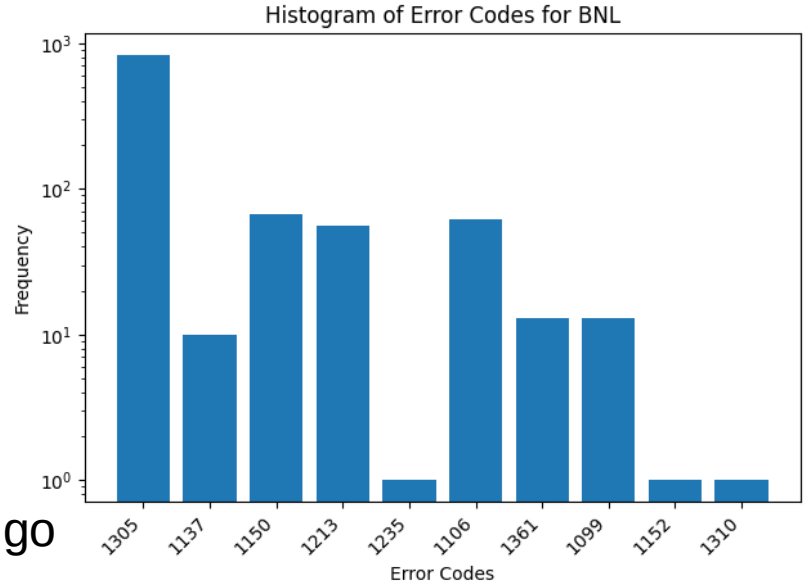
- Dispatcher takes two input: prioritized job queue & platform resource details
- Jobs are then assigned to CPUs at various sites
- Initial “simple” algorithm is used to match jobs to resources
  - Sites ranked according to quality of worker nodes
    - E.g. storage, cores, speed, computational capacity, etc
  - For each job, an optimal CPU is found by looking at the CPU load on the queues (search depth limited to 20)
    - CPU quality reduces as load on it increases
  - Round-robin strategy used once 50% of site CPUs are allocated
- Dispatched jobs sent to SimGrid engine for execution



Plug-in Architecture

- There are several different kinds of job allocation algorithms we will use in this project
  - Simple algorithm for initial testing, historic data for validation, more complex algorithms to be developed by REDWOOD Track 1 & Track 4
- Need a flexible design which let's any developer design a job allocation algorithm and plug it in the simulation, without having to change the core code
- Solution: Plug-ins in the form of shared libraries that can be plugged in the simulator at run time
  - (Currently in development)

Simulation of Errors



- 5-10% of all grid jobs fail for different reasons
  - Need to mirror this for the SimGrid simulation to be realistic
- Simulate errors in SimGrid using real data from PanDA
  - Relevant categories depending on how deep the simulation will go
    - Errors on the simulated “worker node” easiest to start with
- Data source from PanDA: (pilot errors only)
  - [last 24h]
  - [historical period]
- When a given job runs, simulation should decide if job should finish or fail
  - Need to know total number of jobs for same time period (add to JSON)
- If failure, draw random error from distribution for the queue the job is running on
  - Depending on which error it is, set job info accordingly (stage-in error e.g., will not spend any CPU, or very little)
  - Also affects simulated data transfers (no output files will be produced when there is a stage-in error, only the job log will be staged out)

Visualization

- Now exporting job and CPU/worker node status to sqlite file that can be read by monitor
  - Provides live info from running simulation
- Would also like to be able to switch on/off worker nodes or even whole sites on the fly from the monitor

Status and Plans (a selection)

- We have tested both WRENCH and SimGrid, and decided on building the simulation around SimGrid
  - Developed a preliminary software description of the ATLAS grid and tested submitting jobs to it and collecting the output statistics
- Need to make things more accurate (CPU, DISK, etc)
  - Currently, the number of CPUs are randomly distributed (to realistic order) [we have the info], disk sizes are preliminary
  - Investigate if disk info is available for all queues, chase it down if missing
- Want to do basic error injection in simulator
- Develop monitoring further
- Move from simple to realistic brokering (longer term goal)
  - Need to discuss/think how exactly to implement this
    - The algorithm used by PanDA is rather complex – for validation purposes it could be good to come as close as possible to the PanDA algorithm
    - Implement in steps of increasing complexity
- Calibration (e.g. connection bandwidths are preliminary) and validation (e.g. using historical data, job completion times)
- Work to be described in upcoming paper(s)

Q&A (1/3)

- Q: "*Can SimGrid scale to ATLAS grid scale?*" [TW]
- A: Potentially yes .. but depending on several factors
  - Model complexity
    - Scalability significantly impacted by level of detail
      - ALL CPUs/cores in ALL queues are simulated (realistic core power, individually connected to routers etc)
      - ALL queues are inter-connected using real ATLAS topology (which includes measured average data transfer bandwidths)
  - Simulation scope
    - Simulating the entire ATLAS grid with all its intricacies may not be necessary
      - E.g. we could start with studying a "mini-grid" of just a few sites, and then scale up gradually
      - In case of severe performance issues, simplify CPUs – only simulate single CPU on a site but with many cores
    - Focus on simulating specific aspects or subsystems of the grid to improve performance
  - SimGrid's "built-in" scalability features
    - Utilize SimGrid's built-in features for handling large-scale simulations, such as "Platform description language" to efficiently model the grid topology and resource characteristics ***programmatically*** – instead of using "Platform XML" [DONE]
  - Computational Resources
    - Unfortunately, SimGrid is not multi-threaded and cannot easily be run in parallel as multiple processes (i.e. running on the grid would be useless)
    - Current strategy is to run simulation with everything maxed out (brute force method), and basically see how/where it slows down later
    - Obviously, the scalability of the simulation will be limited by the machine running the simulation
      - On the other hand, currently 1M jobs only takes 1 ½ h to execute on standard laptop (will of course get slower with added complexity)

Q&A (2/3)

- Q: "*Do we believe a SimGrid model of the ATLAS grid would be a useful testing ground for PanDA decision making algorithms? (We can't try potentially disruptive changes at scale on the real ATLAS grid)*" [TW]
- A: Arguably, a SimGrid model of the ATLAS grid would be a highly valuable testing ground for PanDA decision-making algorithms – SimGrid is a powerful tool, if used correctly..
  - Safe Experimentation
    - SimGrid provides a controlled environment where new algorithms can be tested and evaluated without the risk of disrupting the real ATLAS grid
    - Simulations allow for precise control over variables, enabling us to isolate the impact of specific algorithm changes and systematically investigate their behavior under different conditions – simulation need to be ~accurate
  - Rapid Prototyping
    - SimGrid facilitates rapid prototyping and development of new algorithms
  - Comprehensive Analysis
    - SimGrid can provide detailed performance metrics, such as job completion times, resource utilization, and overall grid efficiency - this can be used to evaluate the performance of different algorithms and identify areas for improvement
    - Simulations can be used to explore a number of scenarios, including peak workloads, resource failures, and unexpected events

Q&A (3/3)

- Q: "*Would a smaller SimGrid model that doesn't stretch scalability be a useful testing ground?*" [TW]
- A: Yes, with a smaller model we concentrate on core logic/behavior of the algorithms without the overhead of managing a massive and complex simulation
  - See "Simulation scope" on slide 12