Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Processing Streaming Data In High Energy Physics Workflows

NATHAN TALLENT, DARREN KERBYSON, MAHANTESH HALAPPANAVAR
MALACHI SCHRAM, KEVIN BARKER, LUIS DE LA TORRE, RYAN FRIESE, JIAN YIN
ERIC STEPHAN, KERSTIN KLEESE VAN DAM

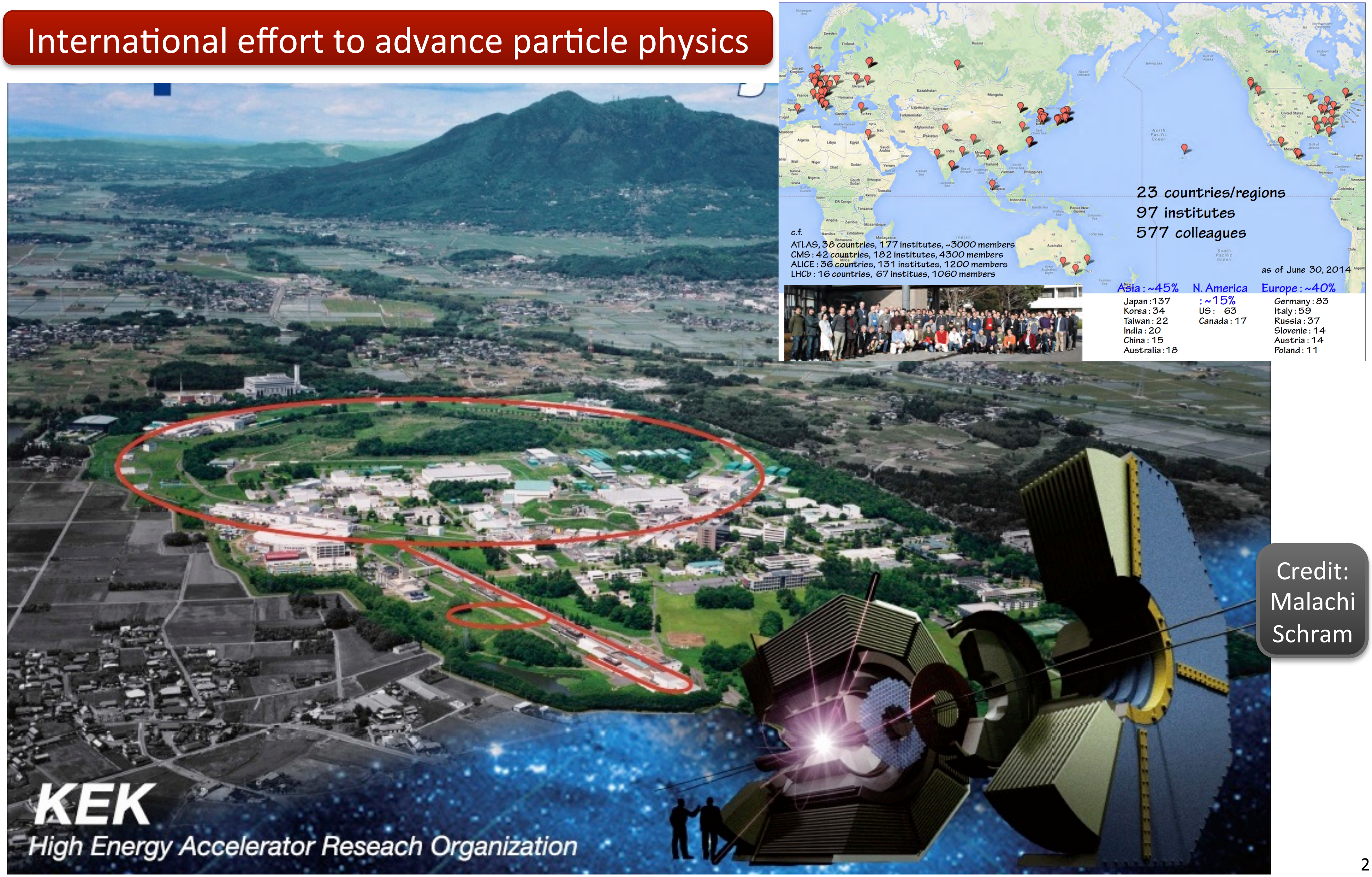
Pacific Northwest National Lab

STREAM '16 Workshop

March 22-23

High Energy Physics: Belle II Analysis Workflow

International effort to advance particle physics



23 countries/regions
97 institutes
577 colleagues

c.f.
ATLAS, 38 countries, 177 institutes, ~3000 members
CMS : 42 countries, 182 institutes, 4300 members
ALICE : 36 countries, 131 institutes, 1200 members
LHCb : 16 countries, 67 instítues, 1060 members

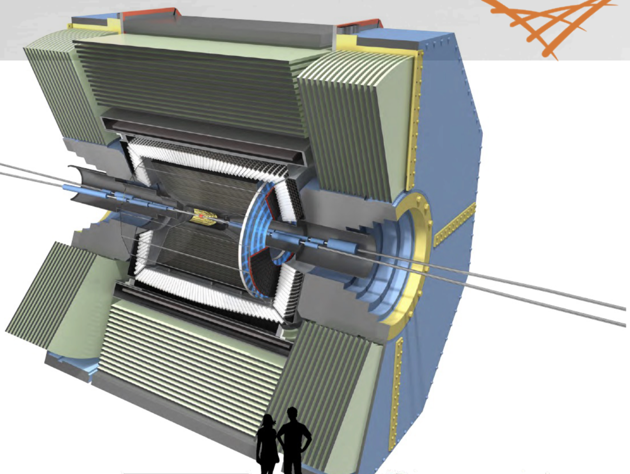
as of June 30, 2014

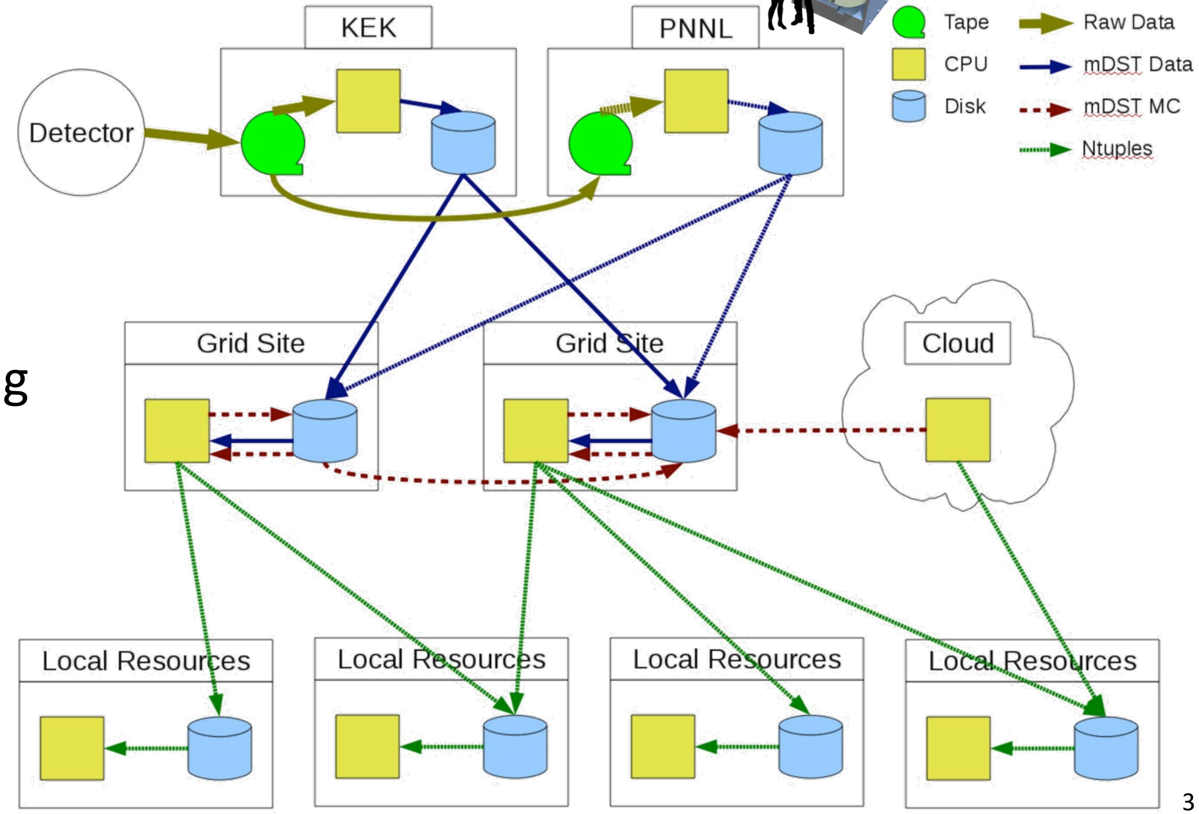
Asia : ~45%	N. America : ~15%	Europe : ~40%
Japan : 137	US : 63	Germany : 83
Korea : 34	Canada : 17	Italy : 59
Taiwan : 22		Russia : 37
India : 20		Slovenie : 14
China : 15		Austria : 14
Australia : 18		Poland : 11

Credit: Malachi Schram

Belle II: Geographically Distributed Analytics

- Belle II Workflow: Extensive data analysis
- Data! 25 PB/year of raw data
  - Stored data expected to reach 350 PB
- Many analysis pipelines run concurrently
  - Normalize raw data
  - Physics analysis
  - Monte Carlo simulations
  - Data storage/archiving





Contention! Many independent data accesses in small window.

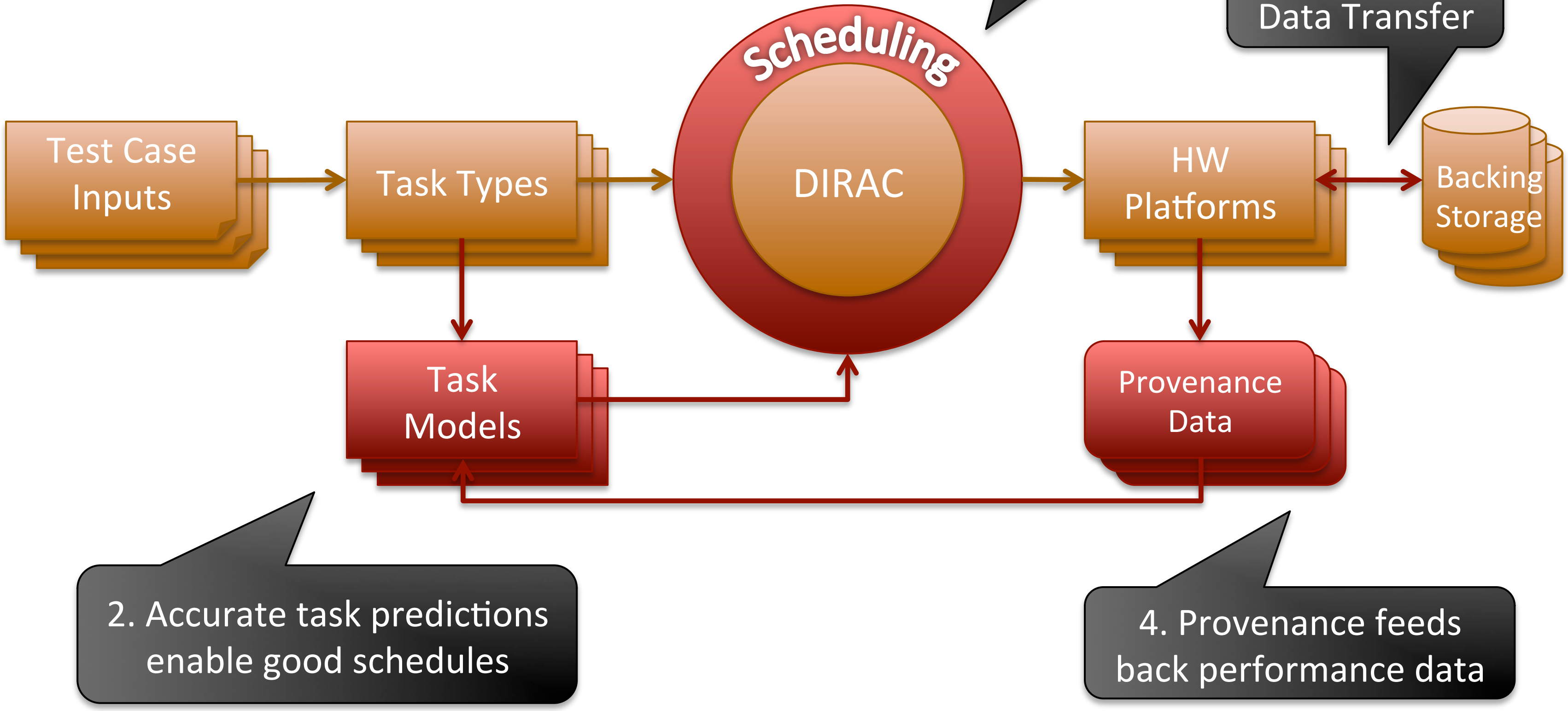
IPPD's 'Enhanced' Belle II Workflow Execution

Challenge

- Mitigate contention
- Consider power/energy

1. Scheduler avoids I/O & network contention

3. Optimized Data Transfer



2. Accurate task predictions enable good schedules

4. Provenance feeds back performance data

IPPD: Integrated End-to-End Performance Prediction and Diagnosis

Hierarchical Scheduler Avoids I/O Contention

Approximate, two-level solution for NP-hard problem

Challenge

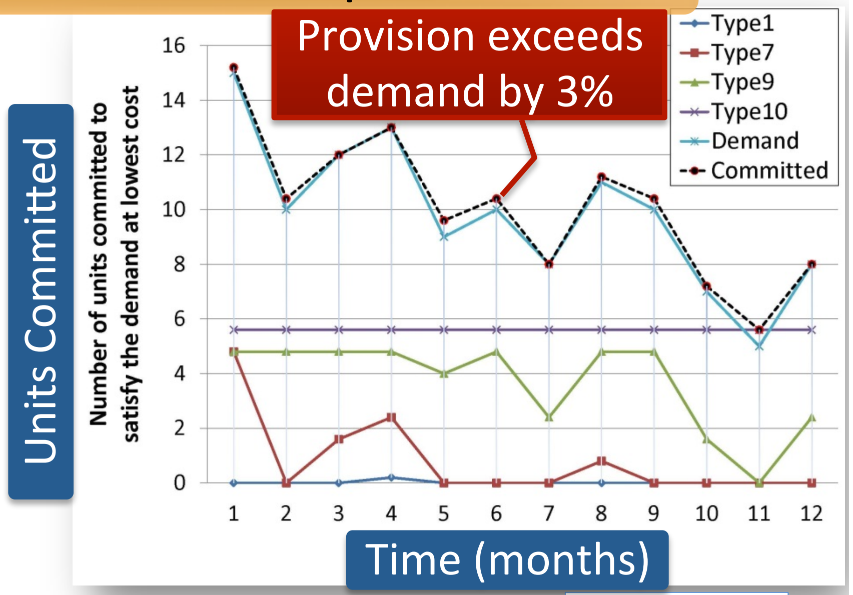
- Demand & supply vary considerably
- Hard to estimate task times
- Congestion dilates execution time

1 Most *cost-efficient subset* of compute resources that meets the tasks' demand

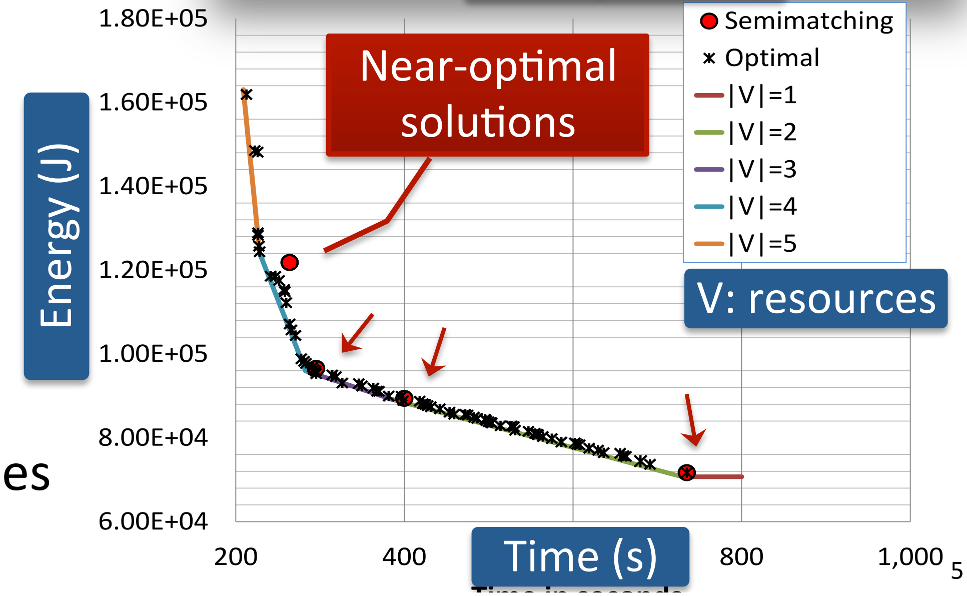
- 🟧 unit commitment (power grid)
- 🟧 mixed int/linear programming

2 Best assignment of tasks to compute resources

- 🟧 bi-objective: energy & time
- 🟧 semi-matching: tasks ↠ resources



Provision exceeds demand by 3%



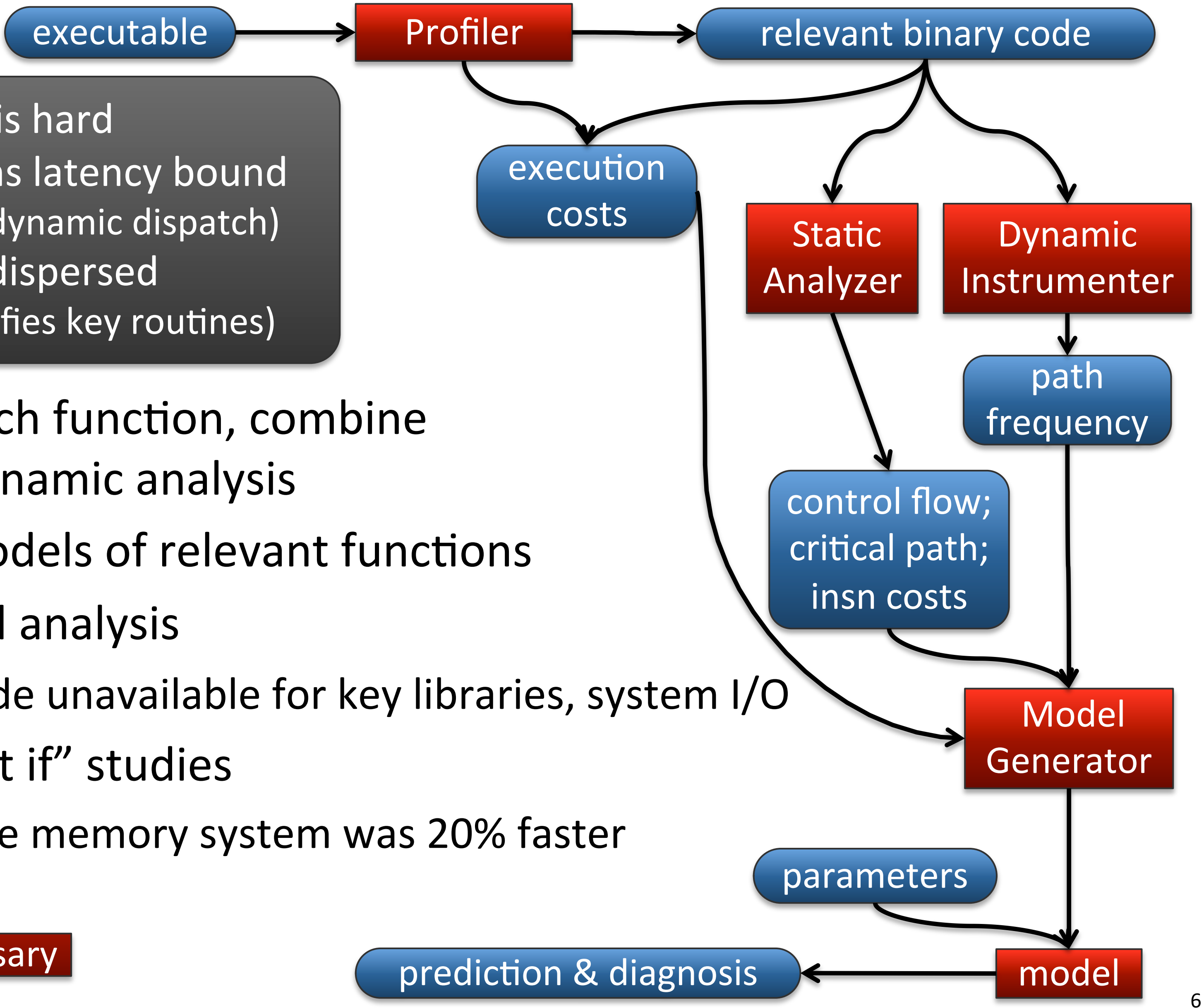
Near-optimal solutions

V: resources

Analytical Modeling Predicts Task Execution Time

Modeling Belle2 is hard

- Most functions latency bound (memory ops; dynamic dispatch)
- Costs widely dispersed (Profiling identifies key routines)



executable

Profiler

relevant binary code

execution costs

Static Analyzer

Dynamic Instrumenter

path frequency

control flow; critical path; insn costs

Model Generator

parameters

model

prediction & diagnosis

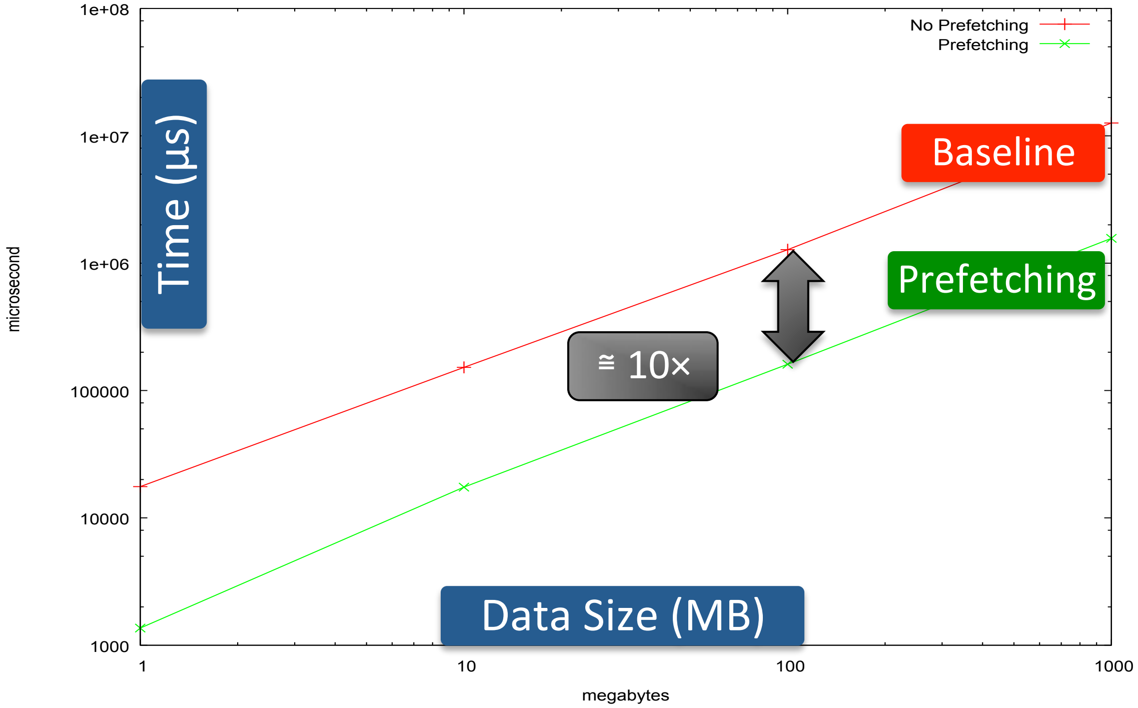
- To model each function, combine static and dynamic analysis
- Compose models of relevant functions
- Binary-based analysis
  - source code unavailable for key libraries, system I/O
- Enable “what if” studies
  - What if the memory system was 20% faster

refine as necessary

Optimize Data Transfer via 'Paced' Prefetching

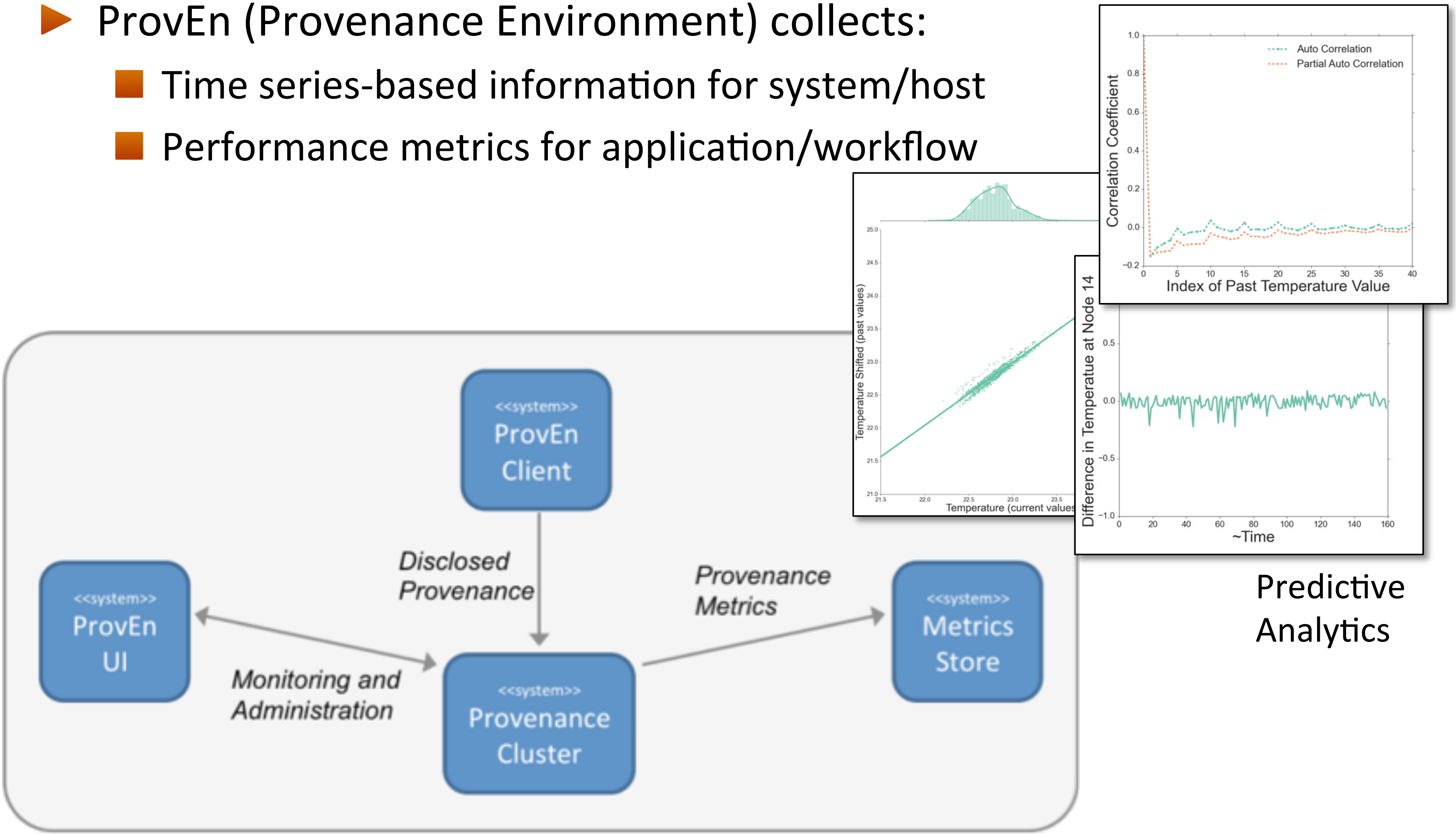
Challenge: I/O Requests Create Blocking Time

- Prefetch data to reduce I/O blocking time
  - Overlap remote data transfer and computation
  - Retrieve only the needed part of a file
  - Split data transfer across multiple internet connections
  - Dynamically adjust given load on each connection
  - Pace I/O request to improve end-to-end performance



Provenance Feeds Back to Scheduler/Modeling

- Provenance delivers execution statistics to scheduler & modeling
- ProvEn (Provenance Environment) collects:
  - Time series-based information for system/host
  - Performance metrics for application/workflow



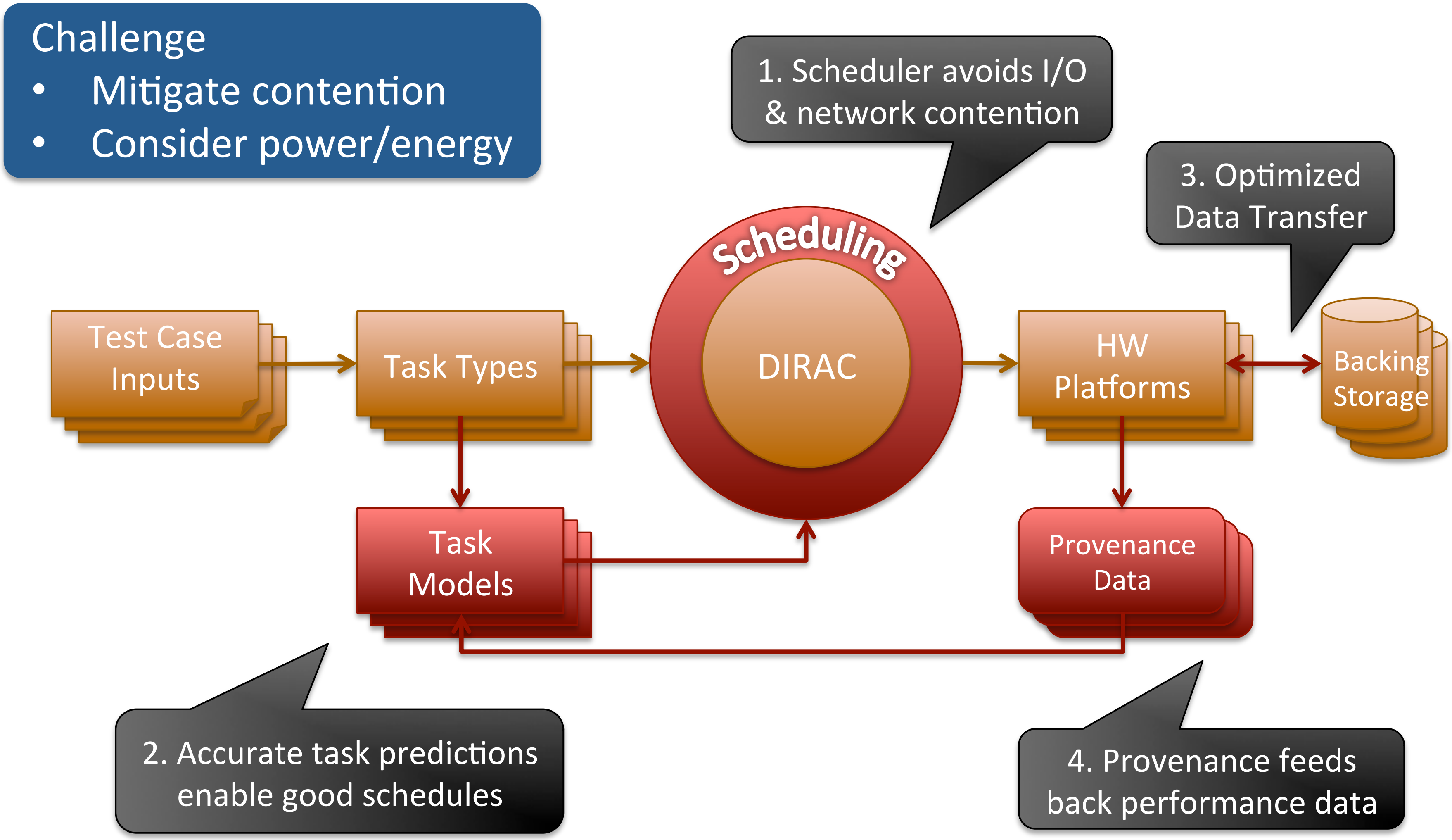
<<system>> ProvEn Client
Disclosed Provenance
<<system>> ProvEn UI
Monitoring and Administration
<<system>> Provenance Cluster
Provenance Metrics
<<system>> Metrics Store
Temperature Shifted (past values)
Temperature (current values)
Difference in Temperatue at Node 14
~Time
Correlation Coefficient
Index of Past Temperature Value
Auto Correlation
Partial Auto Correlation

Predictive Analytics

IPPD's 'Enhanced' Belle II Workflow Execution

Challenge

- Mitigate contention
- Consider power/energy



1. Scheduler avoids I/O & network contention

2. Accurate task predictions enable good schedules

3. Optimized Data Transfer

4. Provenance feeds back performance data

IPPD: Integrated End-to-End Performance Prediction and Diagnosis