CharmROSS

Empowering PDES with an Adaptive Runtime System

UIUC: Eric Mikida, Nikhil Jain, Laxmikant Kale
RPI: Elsa Gonsiorowski, Chris Carothers
LLNL: Peter Barnes, David Jefferson

DES - Background

- Discrete Event Simulation
- Logical Processes (LPs) execute events
- Events have virtual timestamps
- Sequential, Conservative, and Optimistic

PDES - Optimistic Simulations

- Events executed speculatively
- Rollback when there's a causality error
- Need to store a history of events
- Need to reclaim event memory
- Synchronize by calculating the GVT

GVT - Global Virtual Time

- Find the smallest timestamp among all PEs
- Must wait for all events to arrive
- It’s impossible to rollback further than GVT
- Commit events and reclaim memory

ROSS - Background

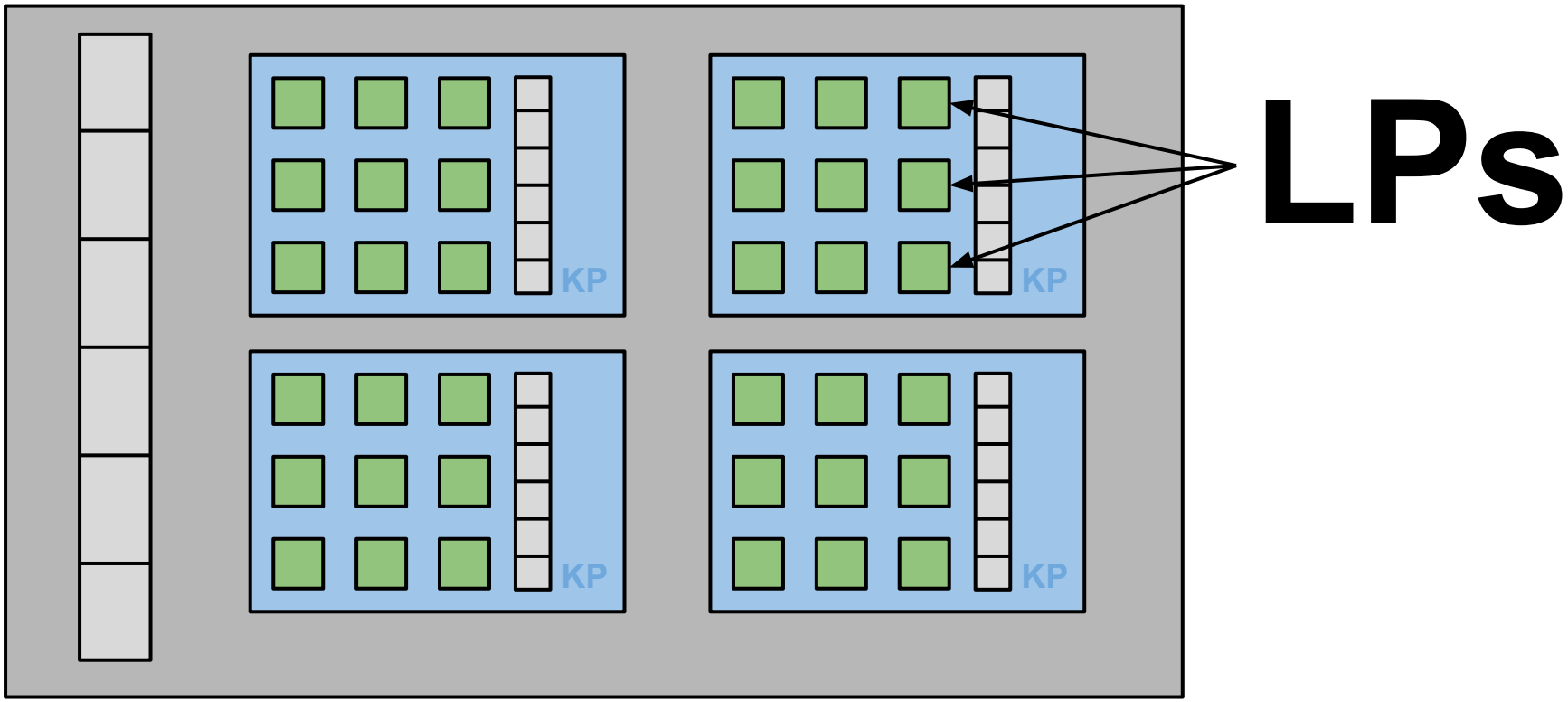
- Developed at RPI
- Written in MPI
- Sequential, Conservative, and Optimistic
- Highly Scalable
- 504B events/s on 120 racks of BG/Q
  - Compared to 5.1B events/s on 1 rack
  - Used the PHOLD benchmark with 10% remote

Motivation and Goals

- Minimal changes to API for model writers
- Achieve similar performance to MPI ROSS
- Add new capabilities
  - Asynchrony (GVT)
  - Load balancing
  - Fault tolerance
  - Checkpoint restart
  - Fine-grain message aggregation

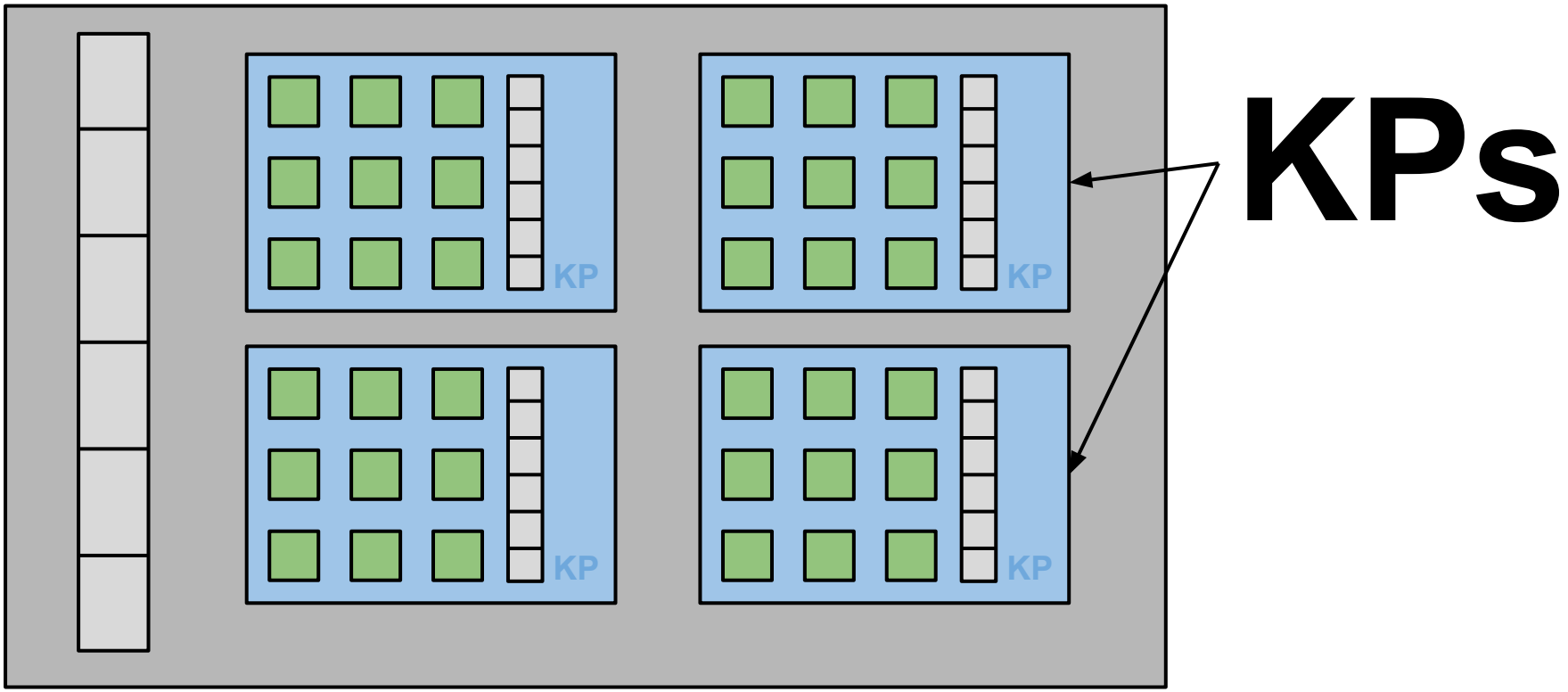
Port Design

MPI ROSS



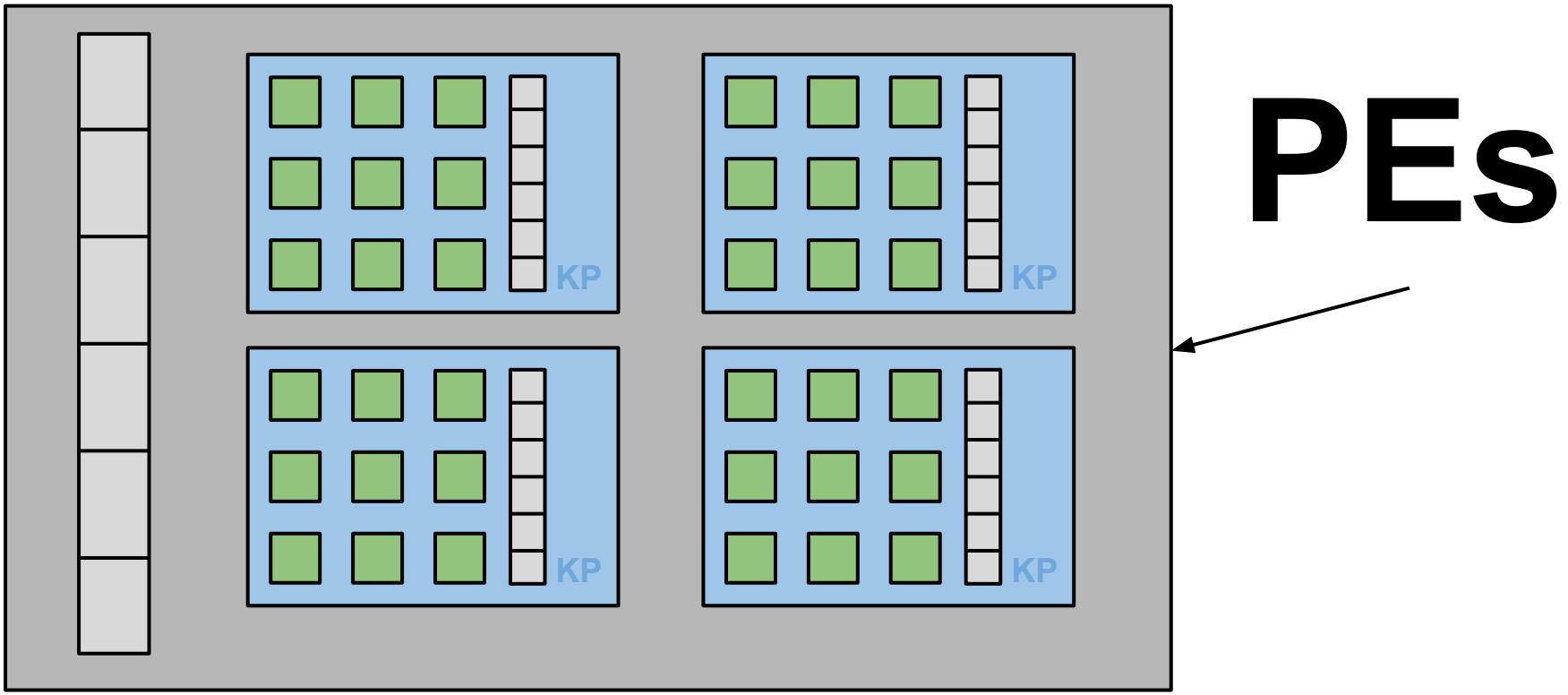
- Behavior defined by the model
- Executes events
- Mapped to KPs

MPI ROSS



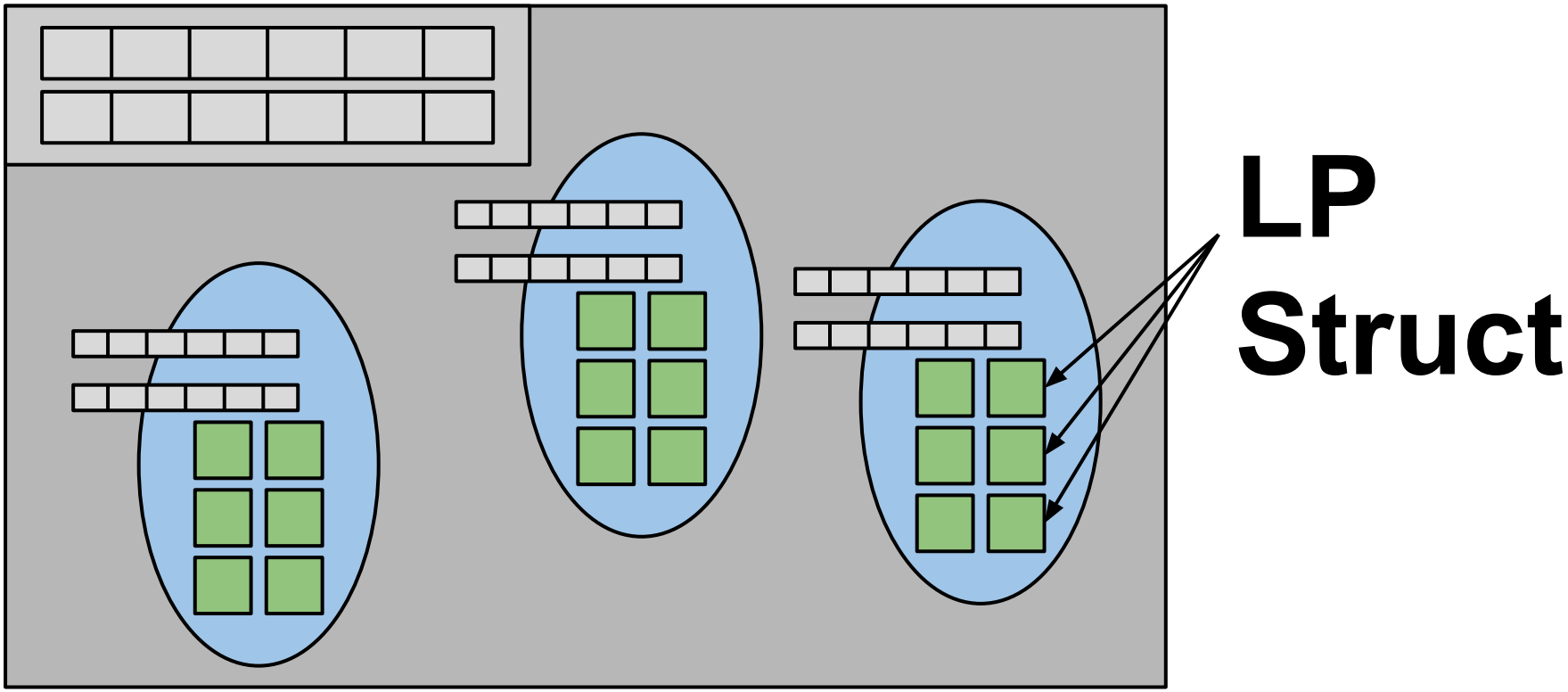
- Stores previous events
- Controls rollbacks and fossil collection

MPI ROSS



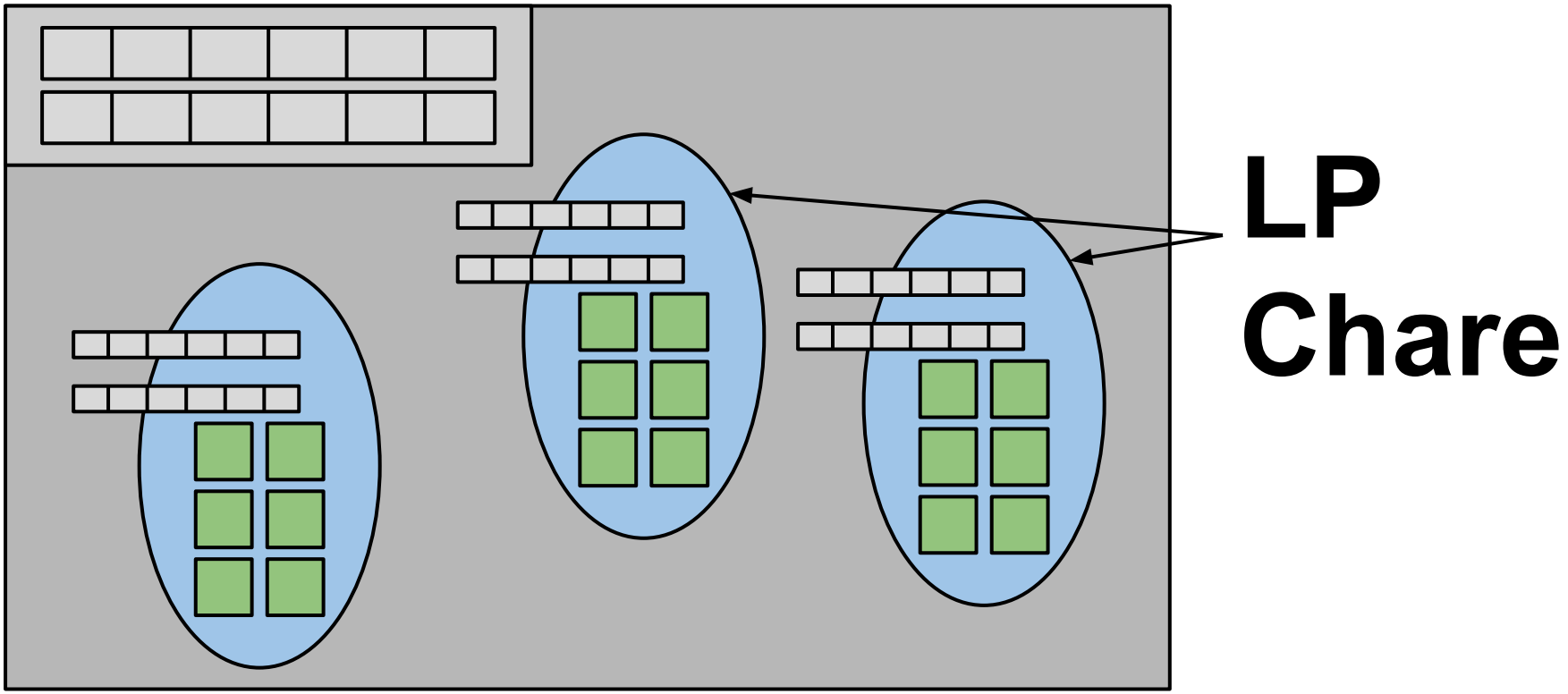
- Stores pending events
- Controls the GVT

CharmROSS



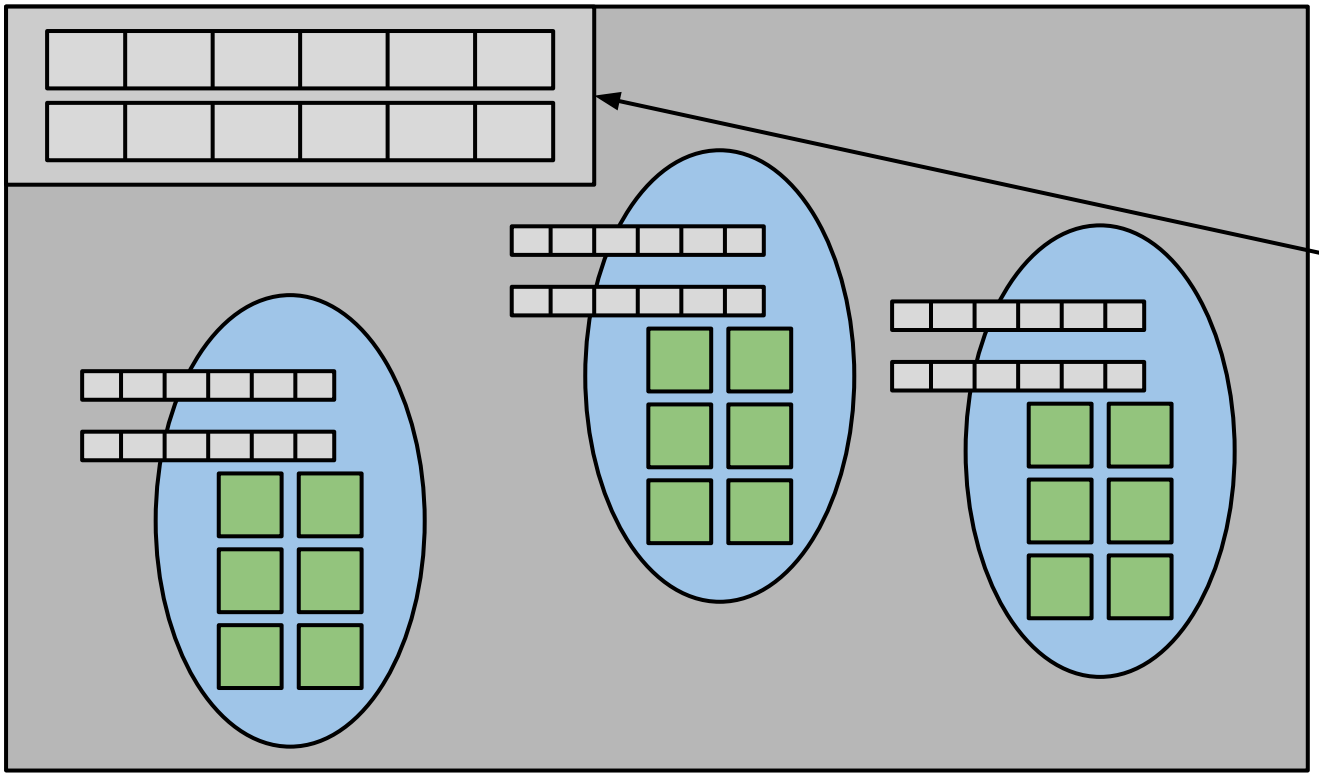
- Identical to LPs in ROSS
- Mapped to migratable chares

CharmROSS



- Combines some KP with some PE
- Holds pending AND past events

CharmROSS



- Now a group chare
- Manages local LP chares
- Controls the scheduler and GVT

Port Status

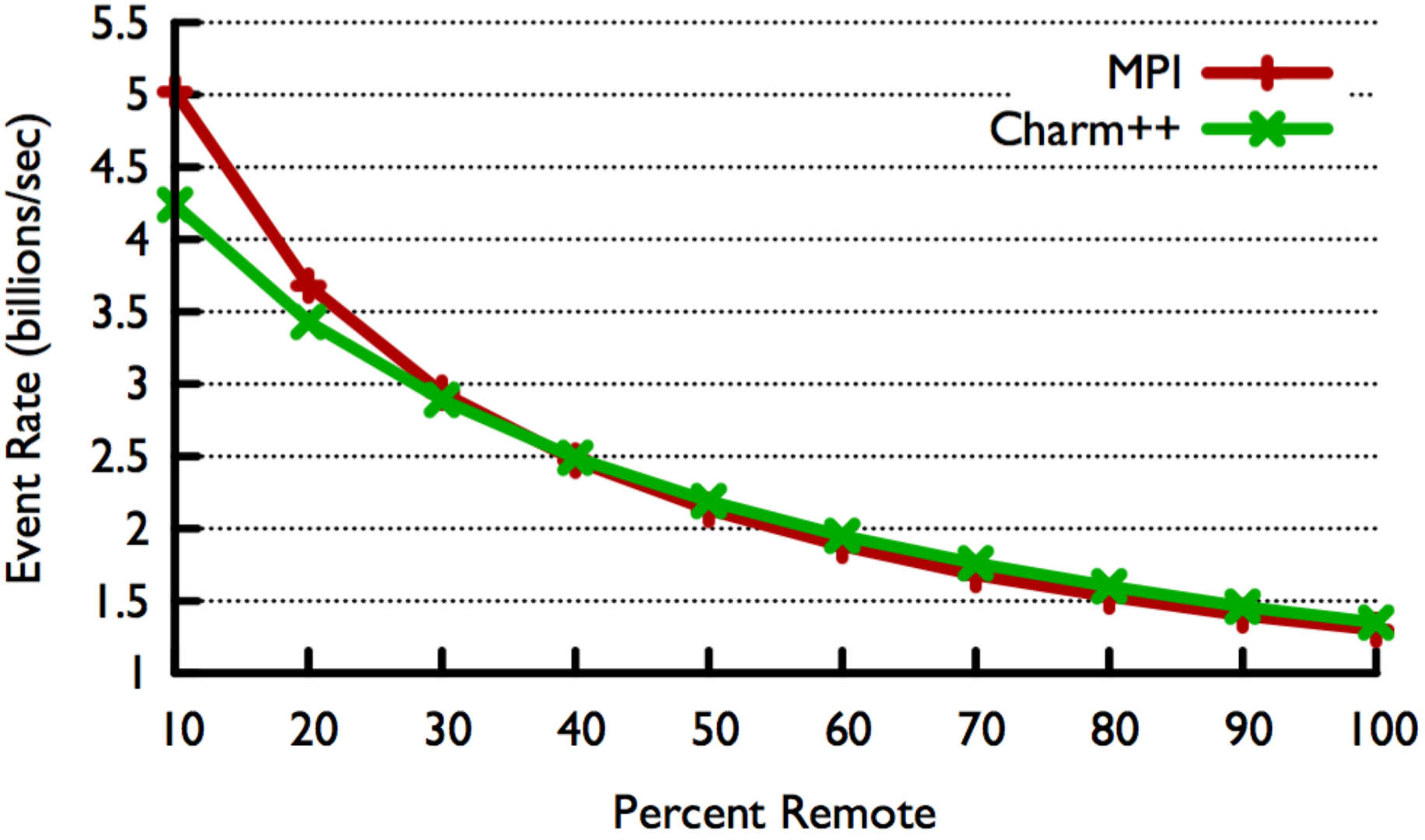
- Sequential, Conservative, Optimistic all work
- Deterministic and consistent with original
- 3 models (PHOLD, PCS, Dragonfly)
- Charm: 4k SLOC, MPI: 8.8k SLOC
- Some extra features implemented

Performance

Initial Performance

- Runs done with PHOLD benchmark
- 1 rack of Vesta (1024 BG/Q nodes)
- 64 threads per node
- No new features included

Varying Remote Communication

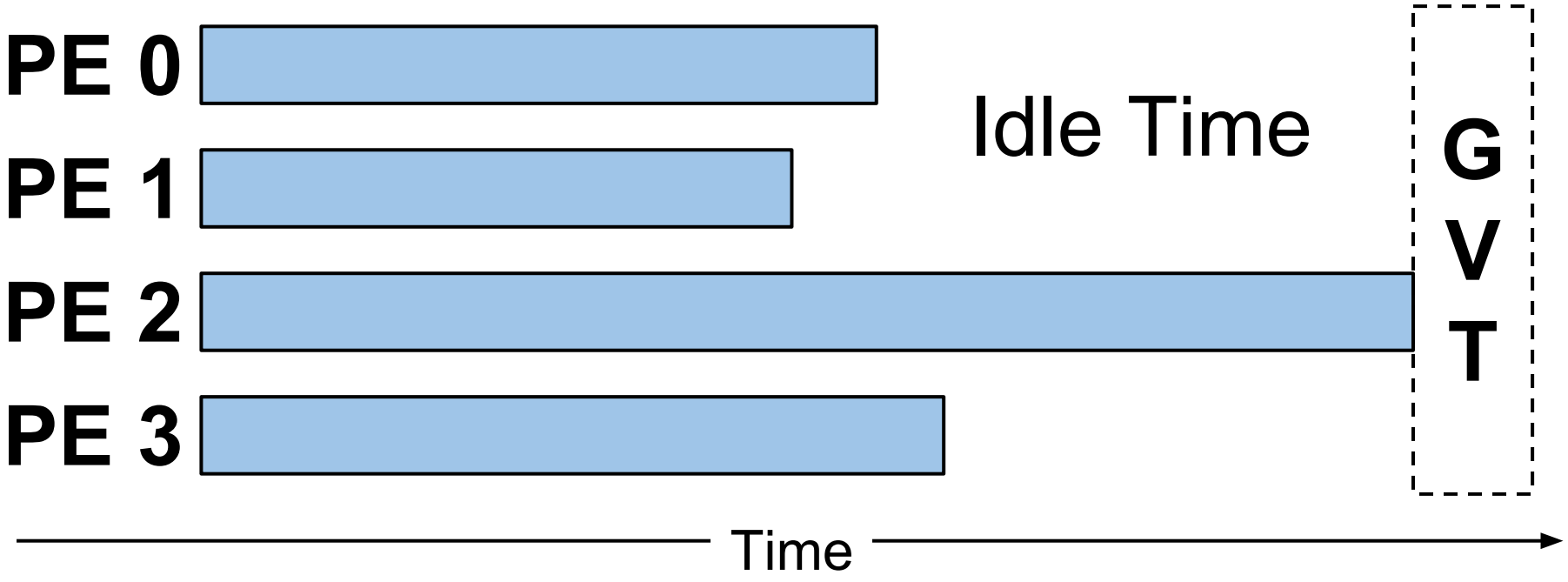


Features

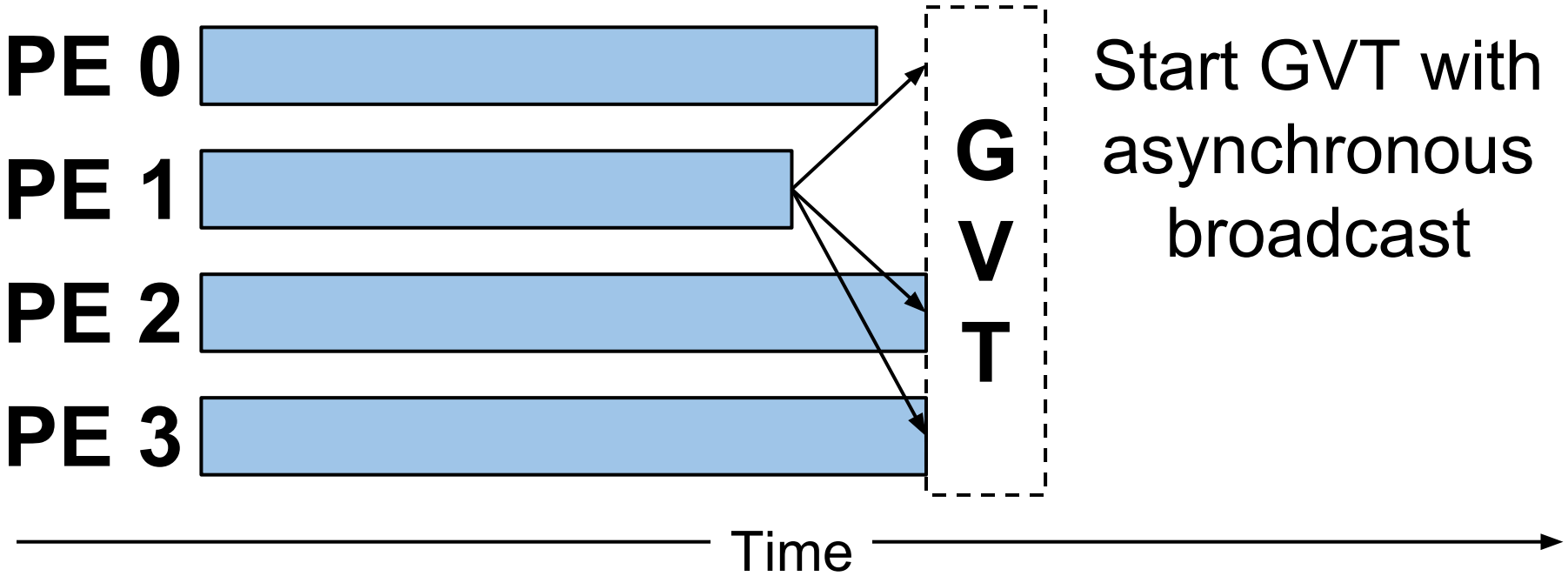
New Features

- GVT Asynchrony
  - Async broadcasts
  - Async reductions
  - Fully async GVT
- Migratability
  - Load balancing
  - Checkpoint/Restart
  - Fault tolerance

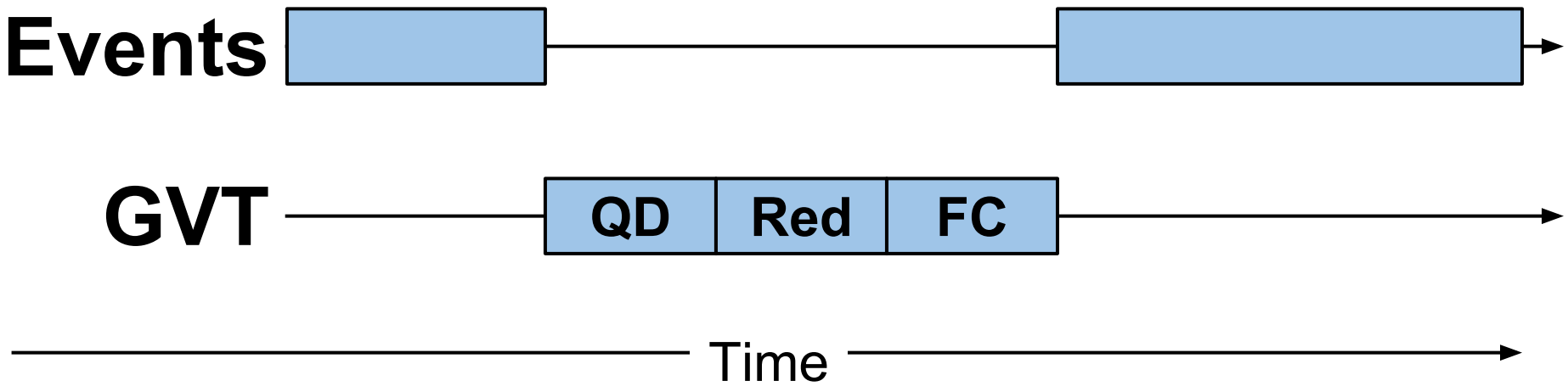
Asynchronous Start



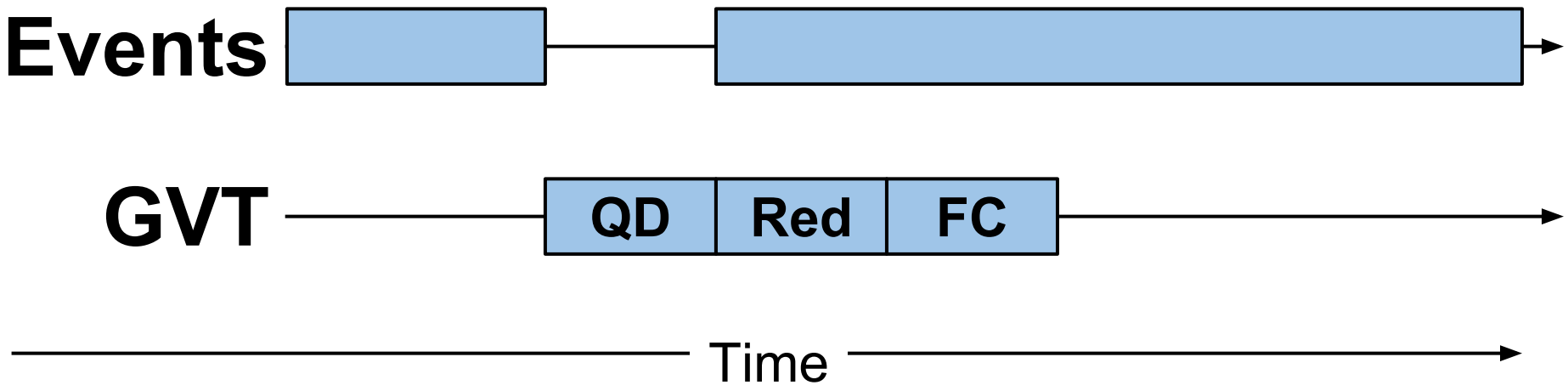
Asynchronous Start



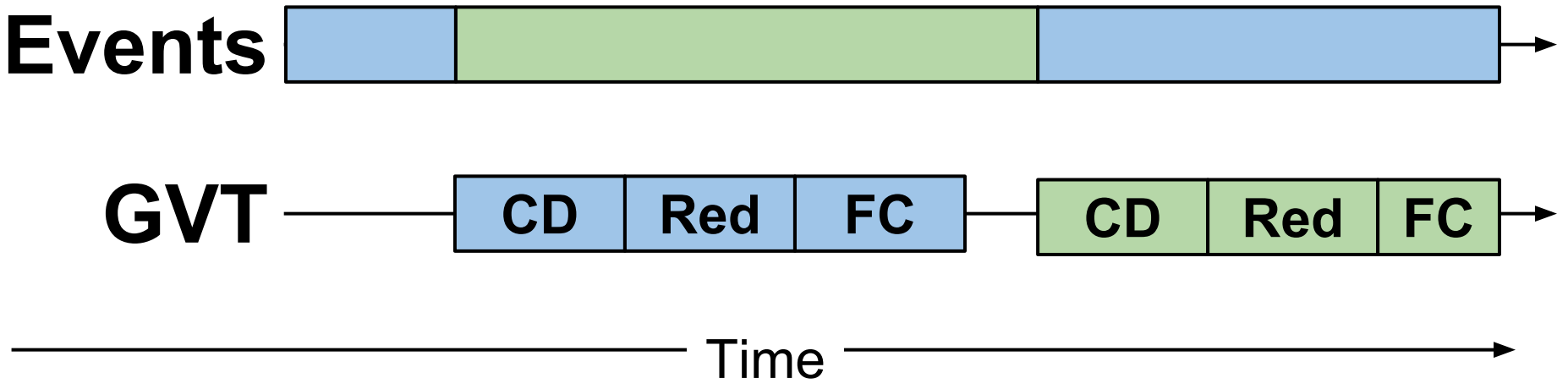
Asynchronous Reductions



Asynchronous Reductions



Fully Asynchronous GVT



Migratability

- LPs are migratable
- Load balancing
- Checkpoint/Restart
- Fault Tolerance

Conclusion

Future Work

- Tuning/optimization of async features
- PDES specific load balancing
- Topological Routing and Aggregation Module