Parallel Discrete Event Simulation (PDES) at ORNL

Presented by

Kalyan S. Perumalla, Ph.D.

Discrete Computing Systems Team
Modeling and Simulation Group
Computational Sciences and Engineering

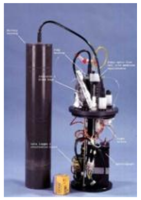
PDES application areas

Many applications in ORNL/DOE mission space and of interest to many other agencies

- High-performance computing system design
  - Application instrumentation, performance tuning, debugging
- Cyber infrastructure simulations
  - Internet protocols, peer-to-peer designs
- Epidemiological simulations
  - Disease spread models, mitigation strategies
- Social dynamics simulations
  - Pre- and post-operations campaigns, foreign policy
- Sensor network simulations
  - Wide area monitoring, situational awareness
- Organizational simulations
  - Command and control, business processes
- Logistics simulations
  - Supply chain processes, contingency analyses







Managed by UT-Battelle
for the U.S. Department of Energy

Implications of discrete event execution on high performance computing

Discrete event execution style is vastly different from most traditional supercomputing-based simulations

Translates to

- Different optimizations
- Different communication patterns
- Different latency needs
- Different bandwidth needs
- Different buffering requirements
- Different scheduling needs
- Different synchronization requirements
- Different flow control schemes

Overall, needs a different runtime

- Qualitatively different runtime infrastructure, designed, built, optimized, and tuned for discrete event applications

Achieved PDES performance at ORNL: A few recent results

- **Parameterized analysis of PDES dynamics**
  - Unique experimental analysis approach and empirical study of discrete event dynamics, on massively parallel platforms
- **New PDES-based virtualized MPI simulator**
  - **µπ System:** Unique, purely PDES-based MPI simulator
  - First-ever achievement of multimillion (virtual) rank MPI execution with synthetic benchmarks, on up to 216K real cores
- **New PDES-based models of epidemic outbreaks**
  - **Epi-RC**: Unique PDES-based model and simulator
  - Significantly surpassed previous best reported scalability of individual-level behaviors in populations of hundreds of millions

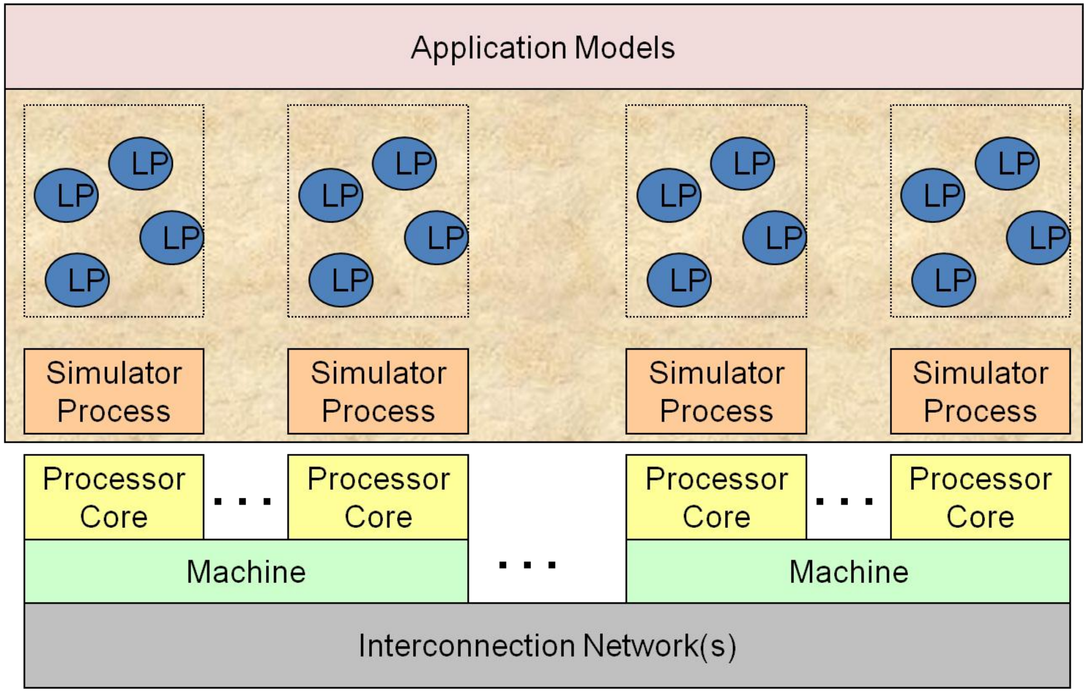
High-performance PDES kernel requirements: Implemented in µsik

- Global time synchronization
  - Total time-stamped ordering of events
  - Paramount for accuracy
- Fast synchronization
  - Scalable, application-independent, time-advance mechanisms
  - Critical for real-time and as-fast-as-possible execution
- Support for fine-grained events
  - Minimal overhead relative to event processing times
  - Application computation is typically low
    - About 2 µs to 50 µs per event
- Conservative, optimistic, and mixed modes
  - Need support for the principal synchronization approaches
  - Useful to choose mode on per-entity basis at initialization
  - Desirable to vary mode dynamically during simulation
- General-purpose API
  - Reusable across multiple applications
  - Accommodates multiple techniques
    - Lookahead, state saving, reverse computation, multicast, etc.

µsik—unique PDES “micro-kernel”

- Unique mixed-mode kernel
- The only scalable mixed-mode kernel in the world
- Supports conservative, optimistic, and mixed modes in a single kernel

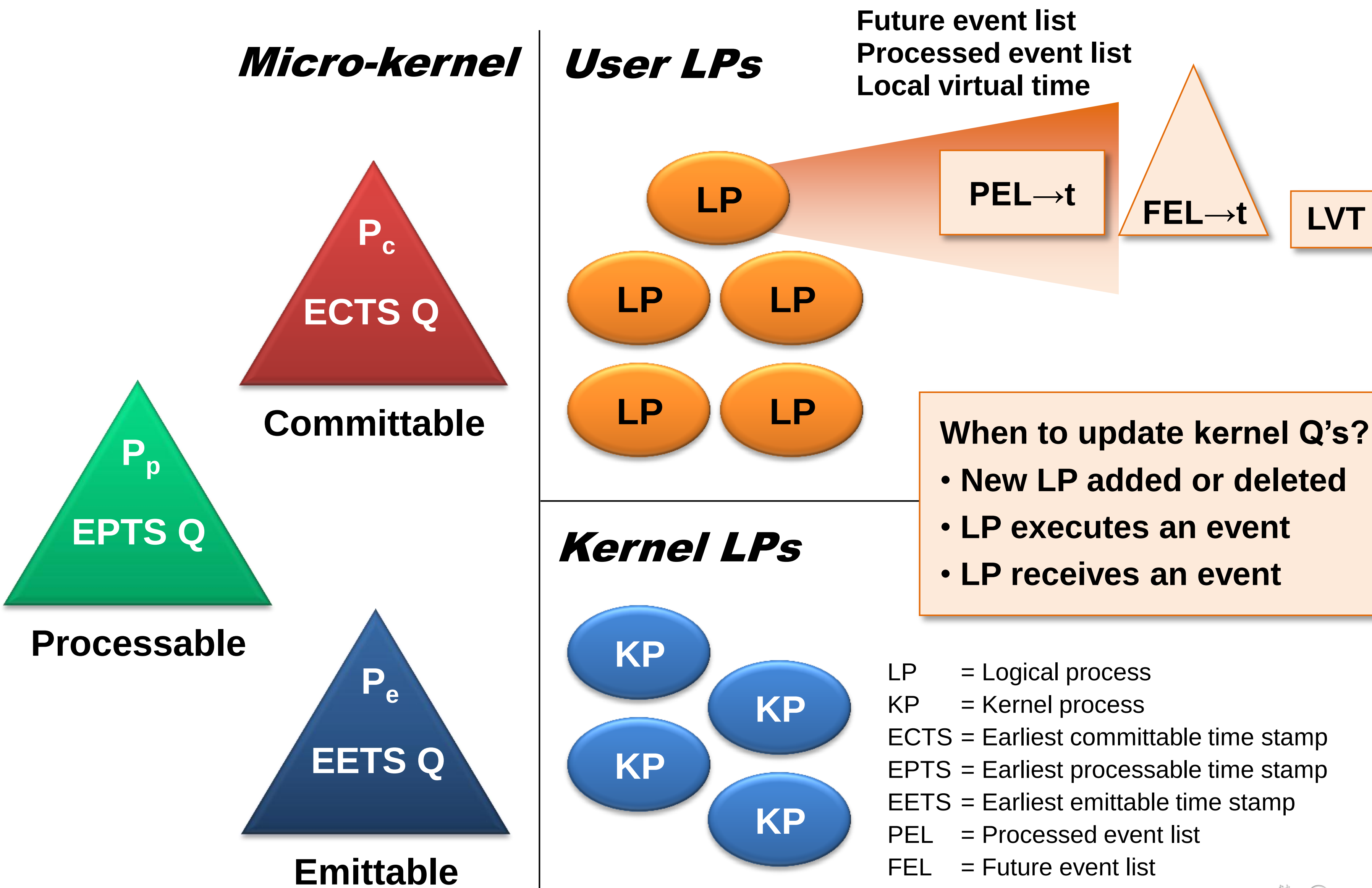
- LP = Logical Process with its own timeline
- Application model entities mapped to LPs exchanging time-stamped events



Used in a variety of applications

- DES-based epidemiological models
- DES-based vehicular traffic models
- DES-based plasma physics models
- DES-based neurological models
- Largest Internet simulations

μsik micro-kernel internals



When to update kernel Q's?

- **New LP added or deleted**
- **LP executes an event**
- **LP receives an event**

LP	= Logical process
KP	= Kernel process
ECTS	= Earliest committable time stamp
EPTS	= Earliest processable time stamp
EETS	= Earliest emittable time stamp
PEL	= Processed event list
FEL	= Future event list
LVT	= Local virtual time

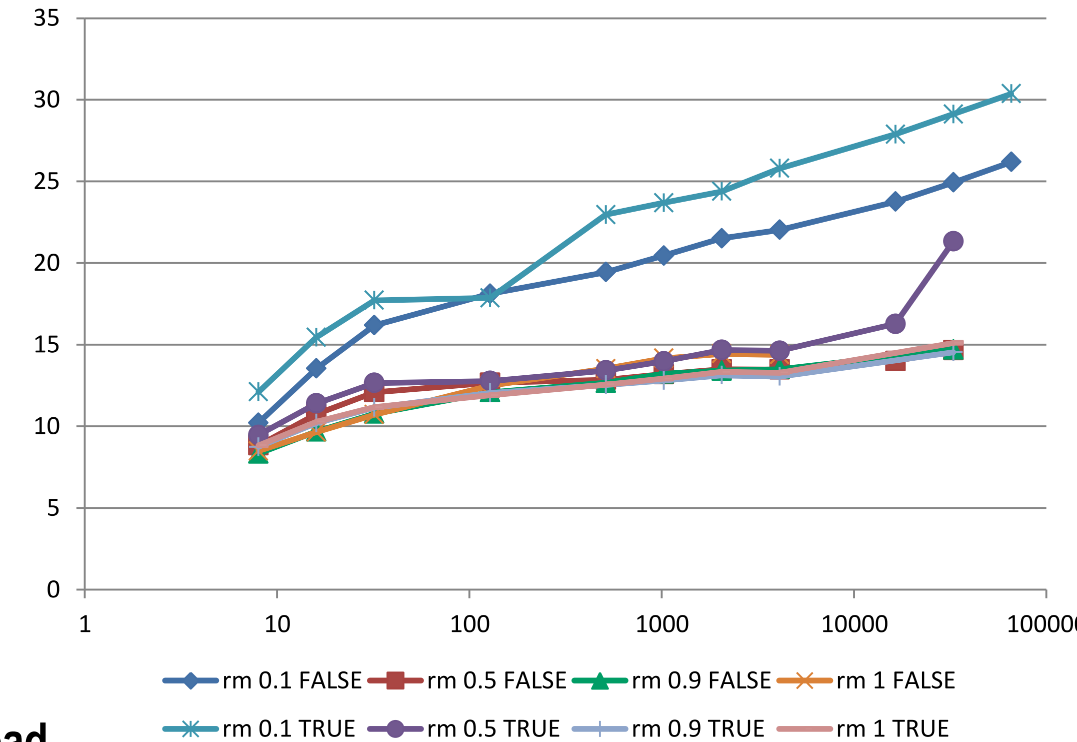
µsik micro-kernel capabilities

- µsik is currently able to support:
  - Lookahead-based conservative and/or optimistic execution
  - Reverse computation–based optimistic execution
  - Checkpointing-based optimistic execution
  - Resilient optimistic execution (zero rollbacks)
    - Constrained, out-of-order execution
    - Preemptive event processing
  - Any combinations of the above
  - Automated, network-throttled flow control
  - User-level event retraction
  - Process-specific limits to optimism
  - Dynamic process addition/deletion
  - Shared and/or distributed memory execution
  - Process-oriented views
- It accommodates the addition of:
  - Synchronized multicast
  - Optimistic dynamic memory allocation
  - Automated load balancing

Analysis of PDES dynamics

- Experimentation with PDES workload spectrum generator for large-scale scenario executions

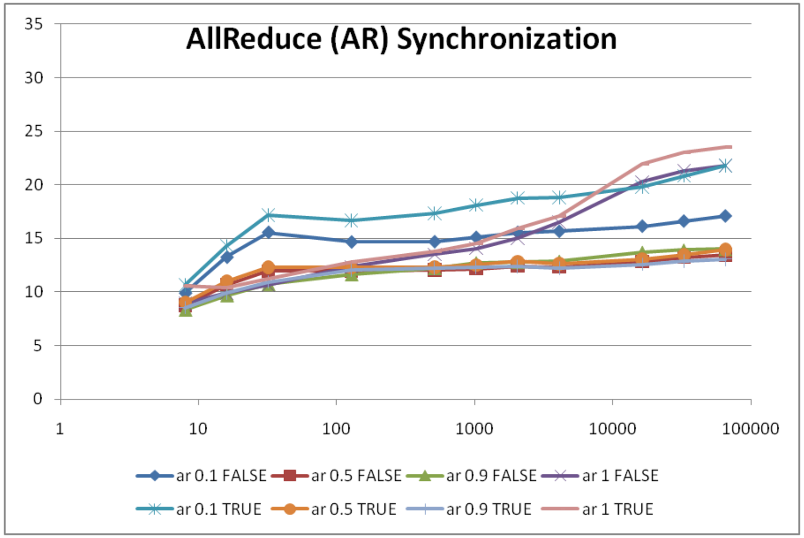
- Exploration of important parameter space
  - (Optimistic, Conservative) ×
  - (Lookahead 0.1–1.0) ×
  - (Allreduce,P2P-Reductions) ×
  - (Inter-rank comm. patterns) ×
  - (8–216K cores)

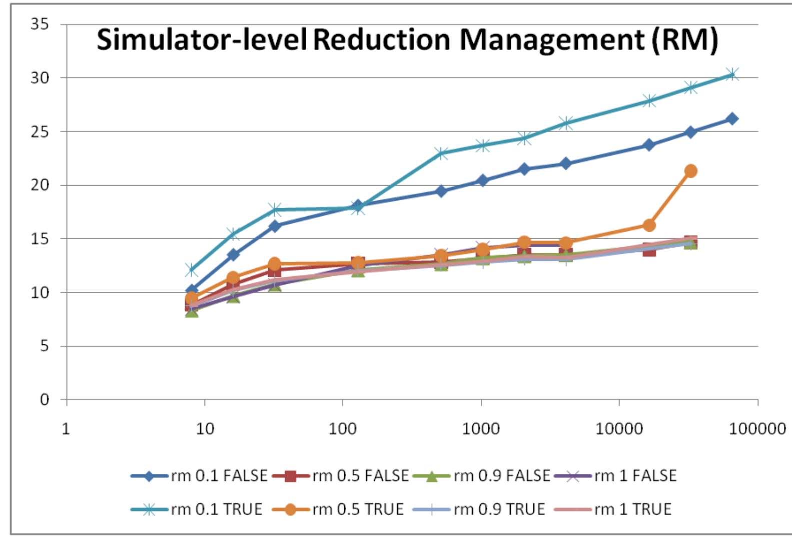


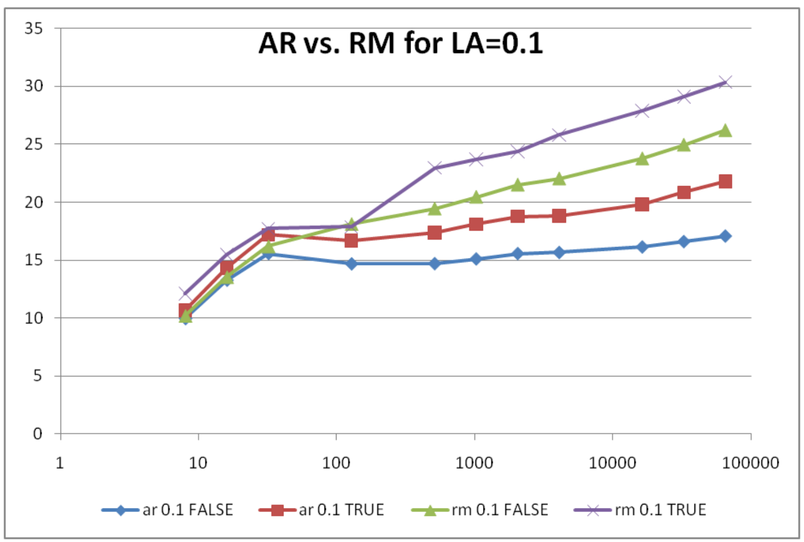
- Discovered important insights into dynamics, e.g.,
  - Allreduce good for low lookahead
  - P2P reductions well-suited for larger lookahead

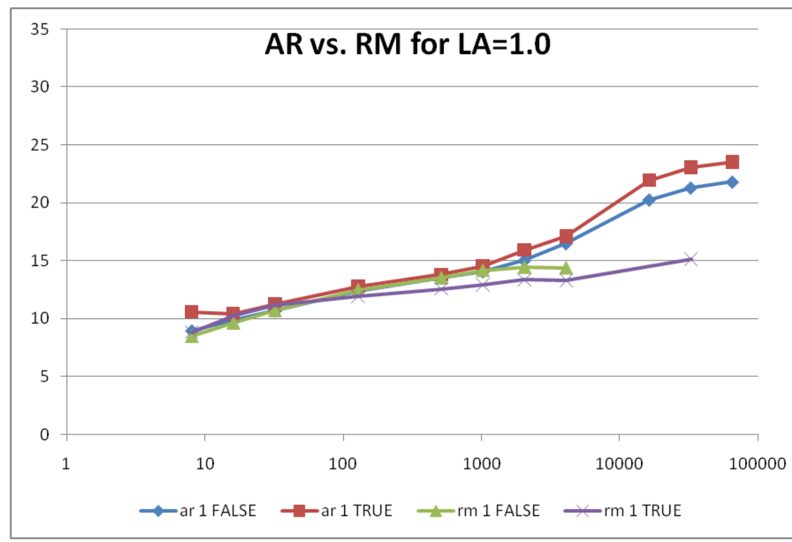
- Preliminary results to appear in IEEE WSC'10

Analysis of PDES dynamics: Event cost vs. number of cores









μπ: PDES-based virtual MPI execution

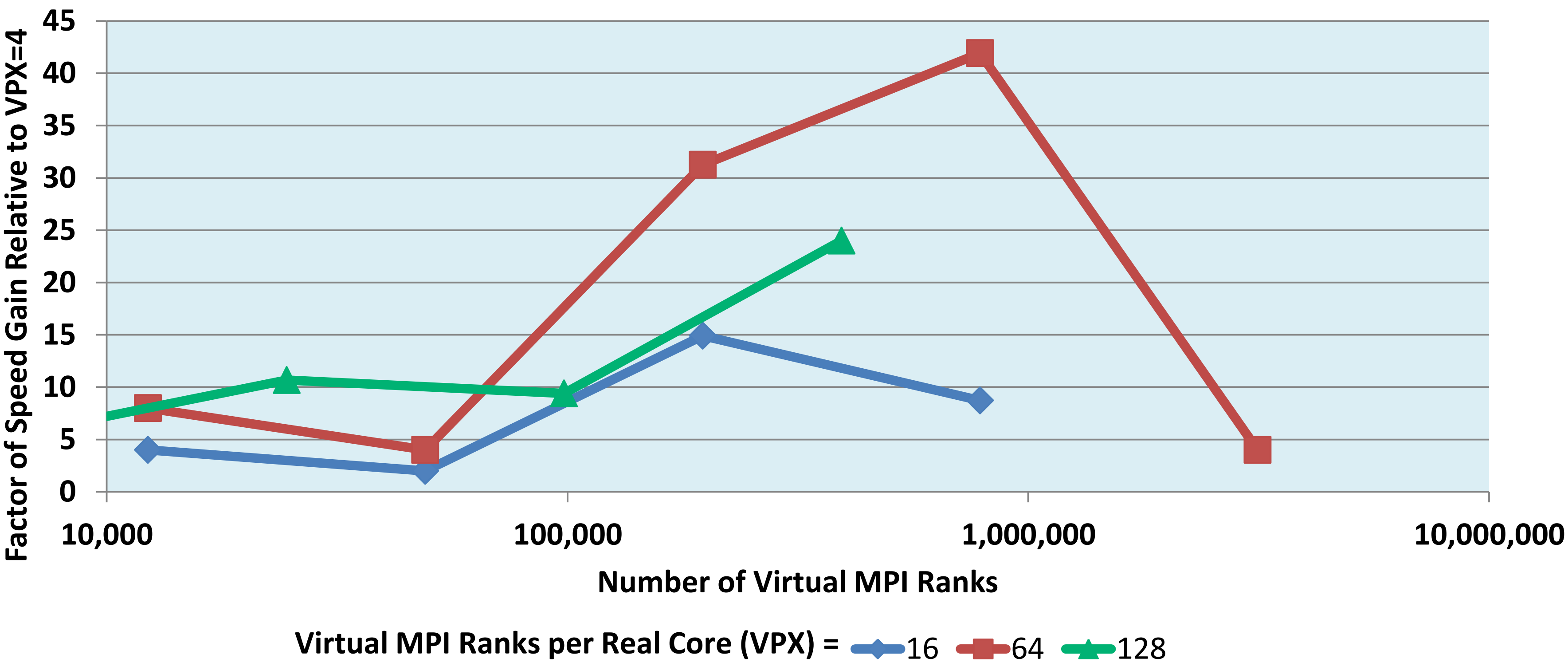
Unique application of parallel discrete event simulation

- Sustaining configurable, controllable, repeatable, and instrumentable execution of MPI programs on larger, envisioned platforms
- Excellent PDES workload, while also being a very useful application in its own right

- Highlights of results
- Very large scale configurations tested
  - 27,648,000 virtual MPI ranks on 216,000 actual cores
  - Full thread-context per virtual rank
- Optimal multiplex-factor empirically determined
  - 64 virtual ranks per real rank
- Showed feasibility to sustain even the most taxing scenarios
  - Zero or low lookahead, corresponding to fast virtual networks

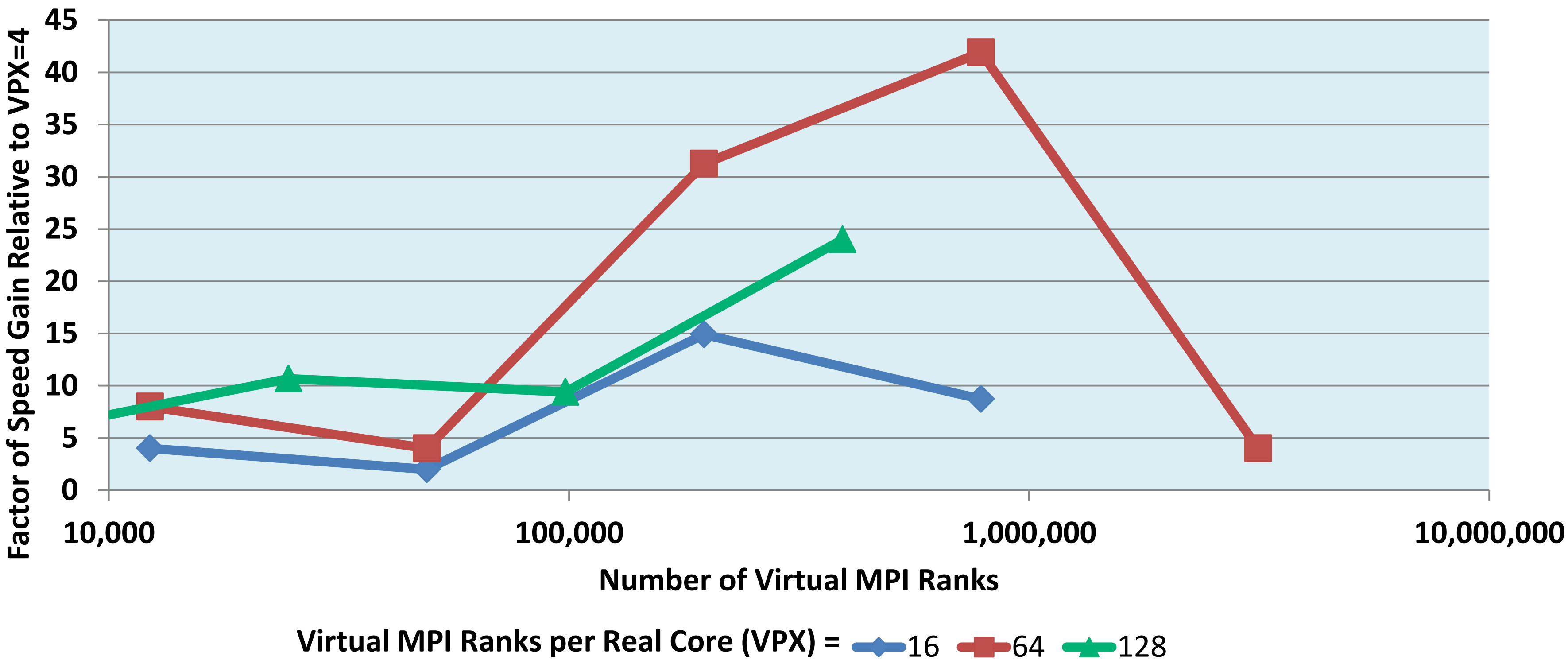
μπ scalability: Runtime cost per event

- Largest (process-oriented) parallel discrete event execution to date
- Weak scaling execution tested on Cray XT4 (up to 16K cores), XT5 (up to 216K cores); also runs on Blue Gene P
- Scalability has been achieved; speed and efficiency are being improved



μπ optimal multiplexing gain: Virtual vs. real ranks

- Uncovered fundamental trade-off between the number of (simulated) virtual ranks desired and the number of actual ranks (cores) that give the best performance to simulate the desired virtual ranks
- Preliminary results documented in ICST SimuTools'10



 Managed by UT-Battelle
for the U.S. Department of Energy

OAK RIDGE National Laboratory

µπ qualitative summary

- The only available simulator for highly scaled MPI runs
  - Suitable for source-available, trace-driven, or modeled applications
- Configurable hardware timing
  - User-specified latencies, bandwidths, arbitrary inter-network models
- Executions repeatable and deterministic
  - Global time-stamped ordering
  - Deterministic timing model
  - Purely discrete event simulation
- Most suitable for applications whose MPI communication may be either trapped, instrumented, or modeled
  - Trapped: on-line, live actual execution
  - Instrumented: off-line trace generation, trace-driven on-line execution
  - Modeled: model-driven computation and MPI communication patterns
- Nearly zero perturbation with unlimited instrumentation

Large-scale high-resolution epidemic outbreak models

Dramatically larger and faster epidemiological simulations than ever before

Discrete Event Reformulation

- More natural representations
- Generalized and flexible structure

Epidemic Outbreak Models

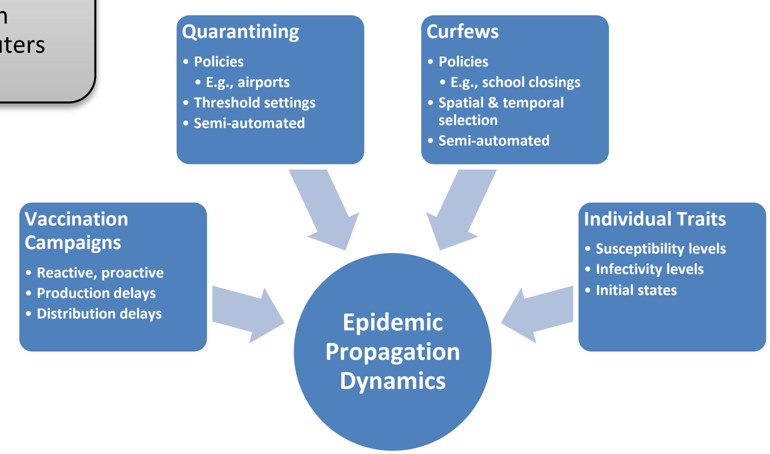
- Reaction-diffusion framework
- High fidelity behaviors

Reversible Parallel Execution

- Enabling reversibility
- Execution on supercomputers

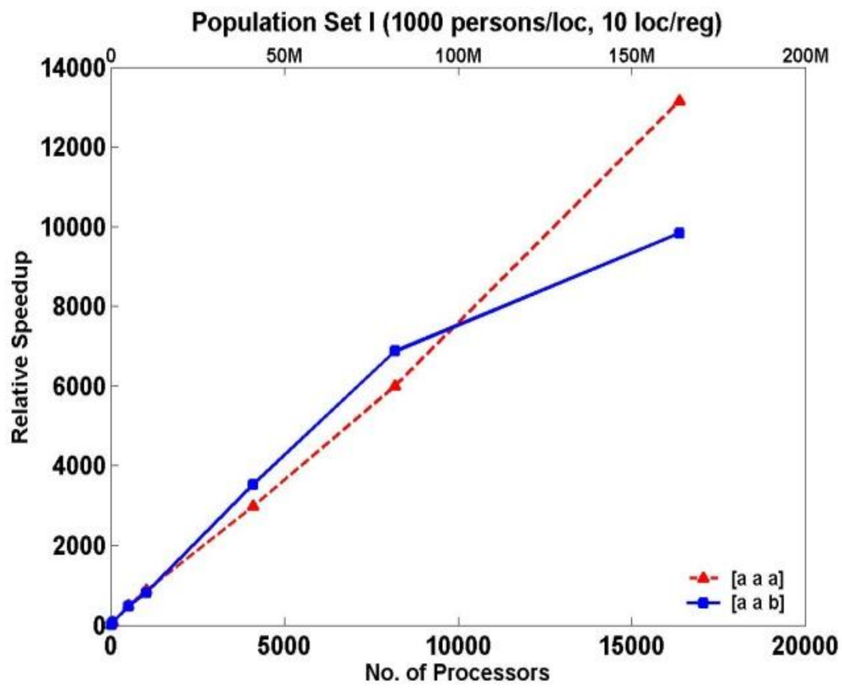
Epi-RC PDES

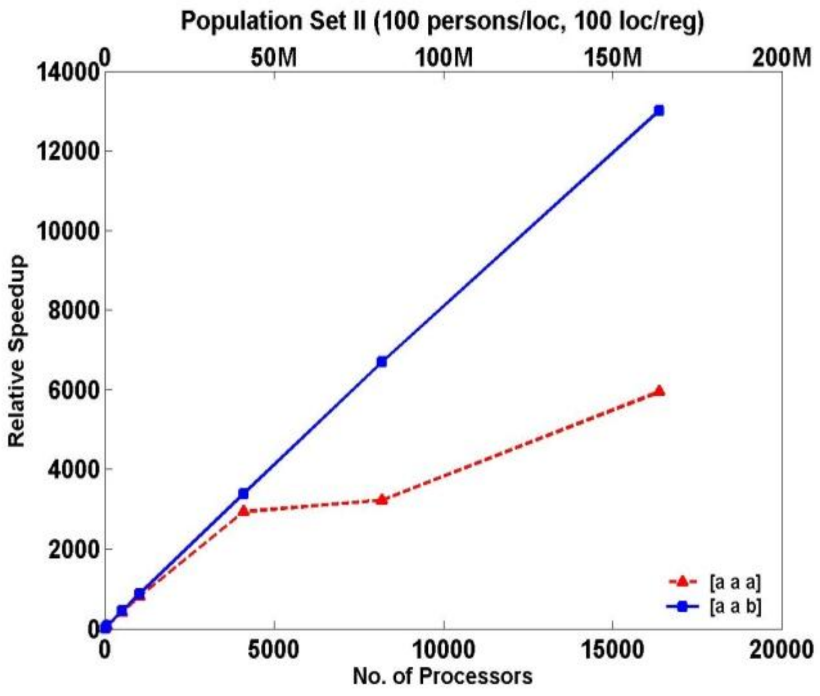
Scaled up to 65,536 cores, billion individuals, much faster than real time



Epi-RC sample runs on Cray XT4

Tag	Type	Set a	Set b
R1	Reaction (Intra)	Level I	Level II
R2	Reaction (Inter)	Infectivity is greater than susceptibility	Susceptibility is greater than infectivity
D	Diffusion (Mobility)	High	Low



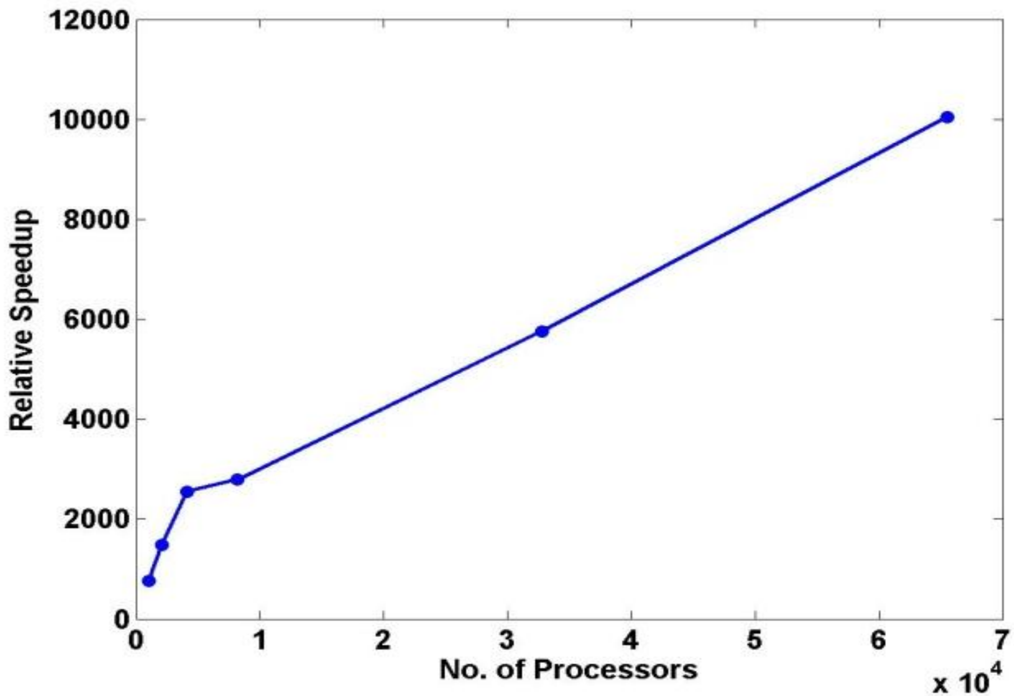


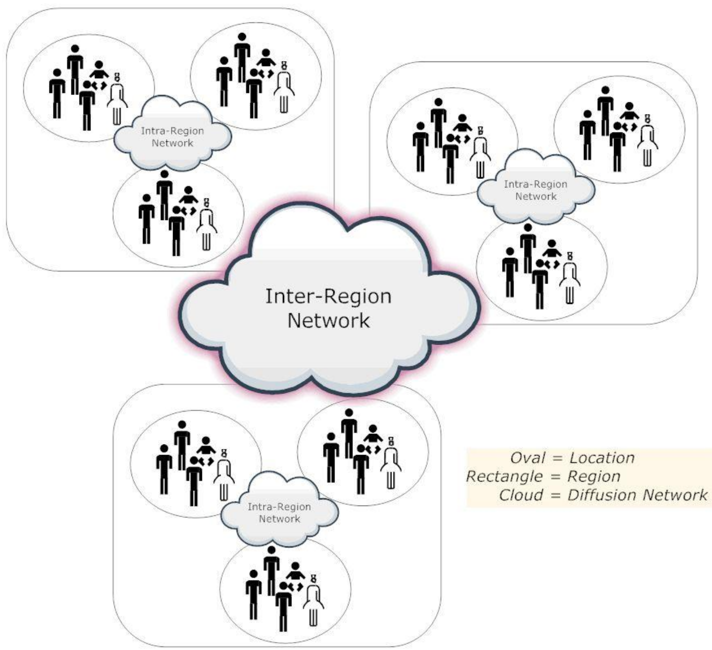
Epi-RC large runs for scalability testing

Speedup of over 10,000 using 65,536 cores

Nearly 1 billion persons exercised in the scenario

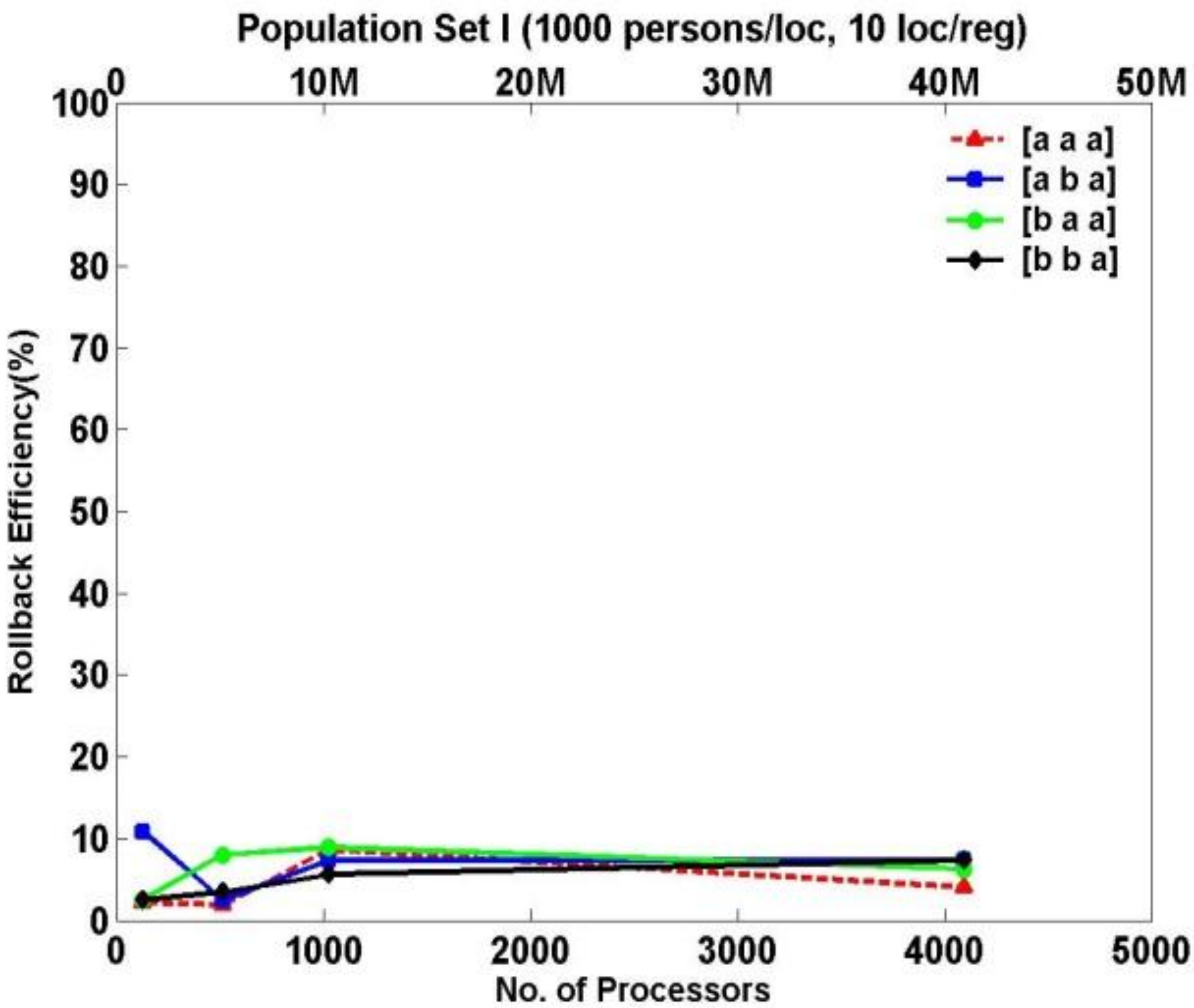
Results reported in PADS'10: double-blind reviewed, best paper finalist





$$p_i = 1 - e^{t \sum_{r \in R} N_r ln(1 - r s_i \rho)}$$

Illustration of Epi-RC rollback dynamics



ORNL's scalable PDES models and µsik highlights

Largest PDES executions to date

- 216,000-core PDES-based execution of multimillion virtual MPI rank scenarios
- 65,536-core execution of new, PDES-based models of high-resolution epidemic outbreaks

Selected articles and presentations

- Perumalla, **"μπ: A Scalable and Transparent System for Simulating MPI Programs,"** ICST SimuTools, 2010
- Perumalla and Seal, **"Reversible Parallel Discrete Event Execution of Large-Scale Epidemic Outbreak Models," IEEE/ACM/SCS PADS, 2010**
- Aaby, Perumalla, and Seal, **"Efficient Simulation of Agent-Based Models on Multi-GPU and Multi-Core Clusters,"** ICST SimuTools, 2010
- Carothers and Perumalla**, "On Deciding between Conservative and Optimistic Approaches on Massively Parallel Platforms,"** IEEE WSC, 2010 (to appear)
- Perumalla and Seal, **"Discrete Event-Based Relaxation for Enabling Concurrency in Naming Game and Other Social Behavioral Models,"** (manuscript)
- Perumalla, **"High-Performance Simulations for Capturing Feedback and Fidelity in Complex Networked Systems,"** SIAM PP10, 2010
- Perumalla and Carothers**, "Compiler-Based Automation Approaches to Reverse Computation,"** Reverse Computation Workshop (in conjunction with IEEE/ACM/SCS PADS'10), 2010

Contact

Kalyan S. Perumalla

Oak Ridge National Laboratory
(865) 241-1315
www.ornl.gov/~2ip
perumallaks@ornl.gov