

Conferences

- International Parallel and Distributed Processing Symposium (IPDPS)
- Symposium on Principles and Practice of Parallel Programming (PPoPP)
- Symposium on Parallelism in Algorithms and Architectures (SPAA)
- International Conference on Distributed Computing Systems (ICDCS)
- ACM/IEEE Supercomputing Conference (SC)
- International Conference on Supercomputing (ICS)
- Symposium on Principles of Distributed Computing (PODC)
- International Conference on High Performance Computing (HiPC)
- International Conference on Parallel and Distributed Systems (ICPADS)

Papers

- Designing efficient SIMD algorithms for direct Connected Component Labeling
- Beyond Human-Level Accuracy: Computational Challenges in Deep Learning
- SEP-Graph: Finding Shortest Execution Paths for Graph Processing under a Hybrid Framework on GPU
- Adaptive Sparse Matrix-Matrix Multiplication on the GPU
- Stretching the capacity of Hardware Transactional Memory in IBM POWER architectures
- Harmonia: A High Throughput B+tree for GPUs
- Engineering a High-Performance GPU B-Tree
- Data-Flow/Dependence Profiling for Structured Transformations
- A Pattern Based Algorithmic Autotuner for Graph Processing on GPUs
- Provably and Practically Efficient Granularity Control
- A Round-Efficient Distributed Betweenness Centrality Algorithm
- Adaptive Sparse Tiling for Sparse Matrix Multiplication
- Efficient Race Detection with Futures
- Checking Linearizability Using Hitting Families
- Bridging the Gap between Deep Learning and Sparse Matrix Format Selection
- Optimizing N-Dimensional, Winograd-Based Convolution for Manycore CPUs
- Cache-Tries: Concurrent Lock-Free Hash Tries with Constant-Time Operations
- Practical Concurrent Traversals in Search Trees
- Communication-Avoiding Minimum Cuts and Connected Components
- LACC: A Linear-Algebraic Algorithm for Finding Connected Components in Distributed Memory
- Shared-memory Exact Minimum Cuts
- Local Distributed Algorithms in Highly Dynamic Networks
- Communication-avoiding CholeskyQR2 for rectangular matrices
- Asynchronous Multigrid Methods
- Practically Efficient Scheduler for Minimizing Average Flow Time of Parallel Jobs
- Reservation Strategies for Stochastic Jobs
- Exploiting Flow Graph of System of ODEs to Accelerate the Simulation of Biologically-Detailed Neural Networks
- Runtime Concurrency Control and Operation Scheduling for High Performance Neural Network Training
- Excavating the Potential of GPU for Accelerating Graph Traversal
- Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems
- Distributed Approximate k-Core Decomposition and Min-Max Edge Orientation: Breaking the Diameter Barrier
- Two Elementary Instructions make Compare-and-Swap
- Accelerating Sequence Alignment to Graphs
- Accurate, Efficient and Scalable Graph Embedding
- Revisiting the I/O Complexity of Fast Matrix Multiplication with Recomputations

- Computation of Matrix Chain Products on Parallel Machines
- Overlapping Communications with Other Communications and its Application to Distributed Dense Matrix Computations
- SunwayLB: Enabling Extreme-Scale Lattice Boltzmann Method Based Computing Fluid Dynamics Simulations on Sunway TaihuLight
- On Optimizing Complex Stencils on GPUs
- Adapting Batch Scheduling to Workload Characteristics: What can we Expect From Online Learning?
- Tight & Simple Load Balancing
- Network Size Estimation in Small-World Networks under Byzantine Faults
- Combining Prefetch Control and Cache Partitioning to Improve Multicore Performance
- Incremental Graph Processing for On-Line Analytics
- Stochastic Gradient Descent on Modern Hardware: Multi-core CPU or GPU? Synchronous or Asynchronous?
- Optimizing Parallel Graph Connectivity Computation via Subgraph Sampling
- Parallel Algorithms through Approximation: b-Edge Cover
- CTA-Aware Prefetching and Scheduling for GPU
- Scheduling Monotone Moldable Jobs in Linear Time
- Scheduling Parallel Tasks under Multiple Resources: List Scheduling vs. Pack Scheduling
- Communication-free Massively Parallel Graph Generation
- Understanding and Modeling Lossy Compression Schemes on HPC Scientific Data
- Hardware Transactional Memory meets Persistent Memory
- Avoiding Synchronization in First-Order Methods for Sparse Convex Optimization
- Blocking Optimization Techniques for Sparse Tensor Computation
- Characterizing Scheduling Delay for Low-latency Data Analytics Workloads
- Runtime Scheduling Policies for Distributed Graph Algorithms
- Local Mixing Time: Distributed Computation and Applications
- Performance and Accuracy Trade-offs of HPC Application Modeling and Simulation
- Efficient Solving of Scan Primitive on Multi-GPU Systems
- GPU-Accelerated Large-Scale Genome Assembly
- A Communication-Avoiding 3D LU Factorization Algorithm for Sparse Matrices
- Overhead-Conscious Format Selection for SpMV-Based Applications
- Tiny Groups Tackle Byzantine Adversaries
- Skueue: A Scalable and Sequentially Consistent Distributed Queue
- Scalable Breadth-First Search on a GPU Cluster
- Linear-Size Hopsets with Small Hopbound, and Constant-Hopbound Hopsets in RNC
- The Arvy Distributed Directory Protocol
- Using Time to Break Symmetry: Universal Deterministic Anonymous Rendezvous
- Multiversion Concurrency with Bounded Delay and Precise Garbage Collection
- Faster Matrix Multiplication via Sparse Decomposition
- Skeap & Seap: Scalable Distributed Priority Queues for constant and arbitrary Priorities
- Persistent Non-Blocking Binary Search Trees Supporting Wait-Free Range Queries
- Efficiency Guarantees for Parallel Incremental Algorithms under Relaxed Schedulers
- Parallel Write-Efficient Geometry Algorithms
- Greedy and Local Ratio Algorithms in the MapReduce Model
- Red-Blue Pebble Game: Complexity of Computing the Trade-Off between Cache Size and Memory Transfers
- BQ: A Lock-Free Queue with Batching
- Online Generalized Caching with Varying Weights and Costs
- An Efficient Wait-free Resizable Hash Table
- On the Distributed Complexity of Large-Scale Graph Computations
- Lock-free Contention Adapting Search Trees

- Theoretically Efficient Parallel Graph Algorithms are Fast and Scalable
- Red-Blue Pebbling Revisited: Near Optimal Parallel Matrix Multiplication
- Predicting Faults in High Performance Computing Systems: An In-Depth Survey of the State-of-the-Practice
- Slim Graph: Practical Lossy Graph Compression for Approximate Graph Processing, Storage, and Analytics
- SW_GROMACS: Accelerate GROMACS on SUNWAY TaihuLight
- Fully Integrated FPGA Molecular Dynamics Simulations
- Replication Is More Efficient Than You Think