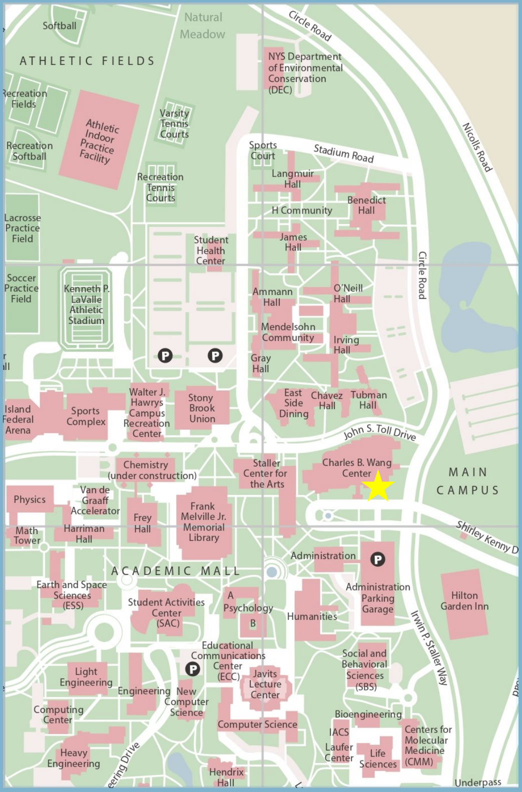

Scan for more info at:

https://www.sigmetrics.org/sigmetrics2025/index.html

Tutorials

Monday, June 9

"Quantum Communications and Networking" , by Saikat Guha (University of Maryland)

"PerfVec: Generalizable Performance Modeling using Learned Program and Architecture Representations", by Lingda Li (Brookhaven National Laboratory), Sairam Sri Vatsavai (Brookhaven National Laboratory), Kuan-Chieh Hsu (Brookhaven National Laboratory)

"Algorithms with Predictions in Queueing: Challenges and Open Problems (Especially for LLMs)", by Michael Mitzenmacher (Harvard) and Rana Shahout (Harvard)

"Distributional Analysis of Stochastic Algorithms", by Qiaomin Xie (University of Wisconsin-Madison), Yudong Chen (University of Wisconsin-Madison)

"Maximizing LLM Throughput in PyTorch: Optimized Pipelines for Modern Deep Learning Workloads", by Davis Wertheimer (IBM, USA)

"Utilizing Underlying Data Statistics in Mitigating Heterogeneity and Client Faults in Federated and Collaborative Learning", by Lili Su (Northeastern University)

"Recent Advances of Reinforcement Learning in Dynamic Games", by Zaiwei Chen (Purdue University, USA) and Kaiqing Zhang (University of Maryland, USA)

"Recent Theoretical Advances in Private Reinforcement Learning", by Xingyu Zhou (Wayne State University, USA)

Workshops

Friday, June 13

AI Crossroads: Systems, Energy, and Applications *(Room: Lecture Hall 2)*

Measurements, Modeling, and Metrics for Carbon-Aware Computing (CarbonMetrics 2025) *(Room: 301)*

Causal Inference Workshop *(Room: 101)*

Learning-augmented Algorithms: Theory and Applications (LATA 2025) *(Room: 201)*

MAthematical performance Modeling and Analysis (MAMA 2025) *(Room: Lecture Hall 1)*

Frontiers in Stochastic Control and Reinforcement Learning (SC&RL) *(Room: 102)*

ACM SIGMETRICS 2025

Keynotes

Sigmetrics Achievement Award
Theater, Tuesday, June 10, 9:15 AM
Devavrat Shah

Scaling AI Computing Sustainably
Theater, Wednesday, June 11, 9:00 AM
Carole-Jean Wu, Director of AI Research, Meta

Responsibly improving advanced AI with privacy-sensitive data
Theater, Thursday, June 12, 9:00 AM
Brendan McMahan, Principal Research Scientist, Google

Invited Talks

Theater, Tuesday, June 10, 1:30 PM - 3:30 PM

Understanding the Host Network
SIGCOMM 2024 Best Student Paper
Rachit Agarwal, Cornell University

Detecting Tiny Performance Regressions at Hyperscale
SOSP 2024 Best Paper & OSDI 2024 Best Paper
Yang Wang, Ohio State University

Tracking, Profiling, and Ad Targeting in the Alexa Echo Smart Speaker Ecosystem
IMC 2023 Best Paper
Umar Iqbal, Washington University in St. Louis

AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration
MLSys 2024 Best Paper Award

Sigmetrics Rising Star Award
Theater, Thursday, June 12, 1:30 AM
Mean-Field Methods for Constrained Systems: Revisiting Load Balancing Under Data Locality
Debankur Mukherjee

Sessions

A: Theater, B: Lecture Hall 1, C: Lecture Hall2

Tuesday, June 10

10:45 AM - 12.:30 PM

Session 1A: Queueing Theory

Steady-State Convergence of the Continuous-Time Routing System with General Distributions in Heavy Traffic
Finite-Time behavior of Erlang-C Model: Mixing Time, Mean Queue Length and Tail Bounds
Improving Multiresource Job Scheduling with Markovian Service Rate Policies
On the Distribution of Sojourn Times in Tandem Queues *(Best Paper Finalist)*

Session 1B: Game Theory

Online Allocation with Multi-Class Arrivals: Group Fairness vs Individual Welfare
Game Theoretic Liquidity Provisioning in Concentrated Liquidity Market Makers *(Best Paper Finalist)*
Two Choice Behavioral Game Dynamics with Myopic-Rational and Herding Players
Allocating Public Goods via Dynamic Max-Min Fairness: Long-Run Behavior and Competitive Equilibria

Session 1C: Security

MUDGUARD: Taming Malicious Majorities in Federated Learning using Privacy-preserving Byzantine-robust Clustering
Application-driven Reexamination of Datacenter Microbursts
VESTA: A Secure and Efficient FHE-based Three-Party Vectorized Evaluation System for Tree Aggregation Models
Confidential VMs Explained: An Empirical Analysis of AMD SEV-SNP and Intel TDX

Wednesday, June 11

10:45 AM - 12:30 PM

Session 2A: Deep Learning

DiskAdapt: Hard Disk Failure Prediction based on Pre-training and Fine-tuning
PROPHET: PRediction Of 5G bandwidtH using Event-driven causal Transformer
Diffusion-Based Generative System Surrogates for Scalable Learning-Driven Optimization in Virtual Playgrounds
FastFlow: Early Yet Robust Network Flow Classification using the Minimal Number of Time-Series Packets

Session 2B: Theory I

Using Lock-Free Design for Throughput-Optimized Cache Eviction
Optimal SSD Management with Predictions
Adversarial Network Optimization under Bandit Feedback: Maximizing Utility in Non-Stationary Multi-Hop Networks *(Best Paper Finalist)*
Reducing Sensor Requirements by Relaxing the Network Metric Dimension

Session 2C: Reliable Systems

The Tale of Errors in Microservices
Quantum Computing in the RAN: Closing Gaps Towards Quantum-based FEC processors
Design and Modeling of a New File Transfer Architecture to Reduce Undetected Errors Evaluated in the FABRIC Testbed
Beaver: A High-Performance and Crash-Consistent File System Cache via PM-DRAM Collaborative Memory Tiering *(Best Paper Finalist)*

1:30 PM - 3:30 PM

Session 3A: Data Centers

Tiered Cloud Routing: Methodology, Latency, and Improvement
Exploring Function Granularity for Serverless Machine Learning Application with GPU Sharing
UniContainer: Unlocking the Potential of Unikernel for Secure and Efficient Containerization
Microns: Connection Subsetting for Microservices in Shared Clusters

Session 3B: Theory II

A Piecewise Lyapunov Analysis of Sub-quadratic SGD: Applications to Robust and Quantile Regression
Optimal Aggregation via Overlay Trees: Delay-MSE Tradeoffs under Failures
The Power of Migrations in Dynamic Bin Packing
Tight bounds for Dynamic Bin Packing with Predictions

Session 3C: Measurement I

Uncovering BGP Action Communities and Community Squatters in the Wild
Beyond App Markets: Demystifying Underground Mobile App Distribution Via Telegram
INT-MC: Low-Overhead In-Band Network-Wide Telemetry Based on Matrix Completion
Beyond Data Points: Regionalizing Crowdsourced Latency Measurements

Thursday, June 12

10:45 AM - 12:30 PM

Session 4A: Online Learning I

Asynchronous Multi-Agent Bandits: Fully Distributed vs. Leader-Coordinated Algorithms
Combinatorial Logistic Bandits *(Best Paper Finalist)*
Online Fair Allocation of Reusable Resources
Universal and Tight Bounds on Counting Errors of Count-Min Sketch with Conservative Updates

Session 4B: Performance

Internet Service Usage and Delivery As Seen From a Residential Network
CHash: A High Cost-Performance Hash Design for CXL-based Disaggregated Memory System
Understanding Intel User Interrupts
A Case Study for Ray Tracing Cores: Performance Insights with Breadth-First Search and Triangle Counting in Graphs

Session 4C: Quantum

Peer-to-Peer Distribution of Graph States Across Spacetime Quantum Networks of Arbitrary Topology
Modeling and Simulating Rydberg Atom Quantum Computers for Hardware-Software Co-design with PachinQo
Optimal Scheduling in a Quantum Switch: Capacity and Throughput Optimality
Quantum Network Optimization: From Optimal Routing to Fair Resource Allocation

2:30 PM - 3:30 PM

Session 5A: Systems I

NetJIT: Bridging the Gap from Traffic Prediction to Preknowledge for Distributed Machine Learning
ScaleOPT: A Scalable Optimal Page Replacement Policy Simulator

Session 5B: Systems II

PipeCo: Pipelining Cold Start of Deep Learning Inference Services on Serverless Platforms
PyGim: An Efficient Graph Neural Network Library for Real Processing-In-Memory Architectures

Session 5C: Blockchains I

CertainSync: Rateless Set Reconciliation with Certainty
The Last Survivor of PoS Pools: Staker's Dilemma

4:00 PM - 5:45 PM

Session 6A: Online learning II

Learning-Augmented Decentralized Online Convex Optimization in Networks
Robust Gittins for Stochastic Scheduling
Learning-Augmented Competitive Algorithms for Spatio-temporal Online Allocation with Deadline Constraints
A Gittins Policy for Optimizing Tail Latency

Session 6B: Measurement II

Revisiting Traffic Splitting for Software Switch in Datacenter
Exploiting Kubernetes Autoscaling for Economic Denial of Sustainability
A Global Perspective on the Past, Present, and Future of Video Streaming over Starlink
ForgetMeNot: Understanding and Modeling the Impact of Forever Chemicals Toward Sustainable Large-Scale Computing

Session 6C: Blockchains II

Phishing Tactics Are Evolving: An Empirical Study of Phishing Contracts on Ethereum
Blockchain Amplification Attack
Piecing Together the Jigsaw Puzzle of Transactions on Heterogeneous Blockchain Networks
Towards Understanding and Analyzing Instant Cryptocurrency Exchanges