Some Science Aspects of Wide-Area Data Transport

Nagi Rao
raons@ornl.gov
Oak Ridge National Laboratory

Mini-Symposium on Data over Distance:
Convergence of Networking, Storage, Transport, and Software Frameworks
July 19, 2018
Hanover, MD

Sponsored by
U.S. Department of Energy
U.S. Department of Defense

Outline

- **Background**
- **Through Profiles of Infrastructures**
  - **Memory and File transfers**
  - **Convexity and Utilization**
- **Profile Estimation: Machine Learning**
  - **Generalization**
- **Cyber-Physical Aspects: Game Theory**
  - **Extension using LNet**
- **Looking Into Future**

Collaborators: S. Sen, Q. Liu (ORNL); Foster, Z. Liu, R. Kettimuthu (ANL); D. Katramatos (BNL), B. Settleymyer, H. B. Chen (LANL); D. Towsley, G. Vardoyan (UMASS); F. He (Texas A&M); J. Zhuang (UBuffalo); C.Y.T Ma (Hong Kong); D. Yau (Singapore)

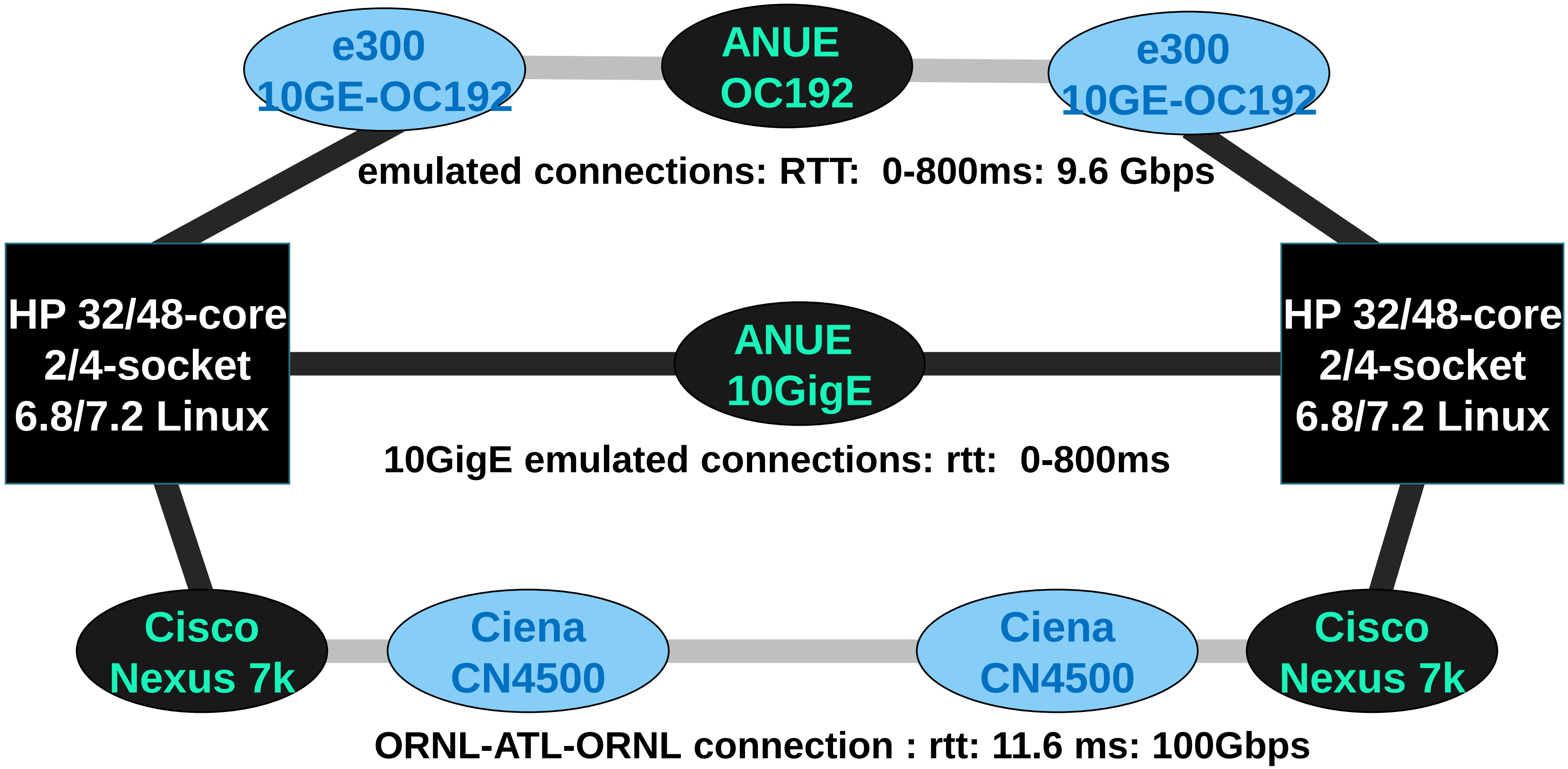
Background

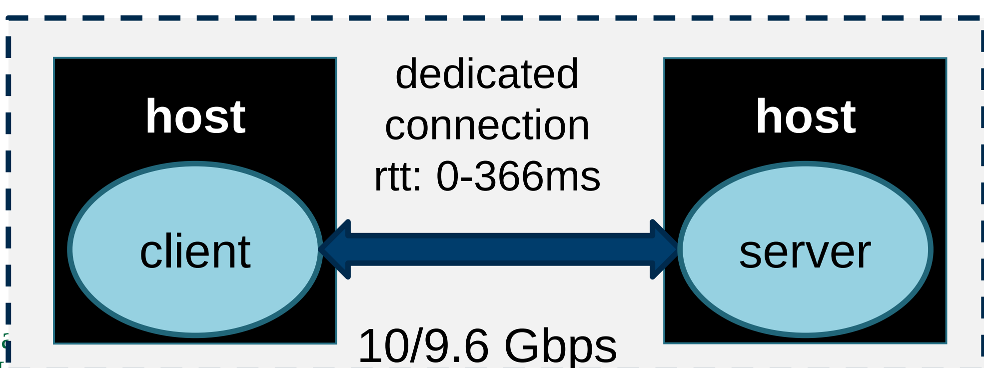
Big-data science and commercial data transport across networks

- Science codes on supercomputers generate large data sets to be transferred to remote storage sites for archival and post-analysis
- Science facilities generate large datasets to be transported to remote supercomputing centers
  - Spallation Neutron Sources at Oak Ridge National Laboratory
- Commercial big data and distributed information systems
  - Google B4 SDN dedicated networks

- **Dedicated Connections**
  - Increasing deployments and availability
    - DOE OSCARS. Google B4
  - Desirable features: dedicated capacity and low loss rates
  - Expectations for transport methods: Simple and predictable flow dynamics
  - Surprisingly, show much more complex profiles and dynamics
    - concave-convex profile vs. convex profile from literature
    - rich dynamics lead to lower performance

ORNL Testbed : Emulated and Physical Connections



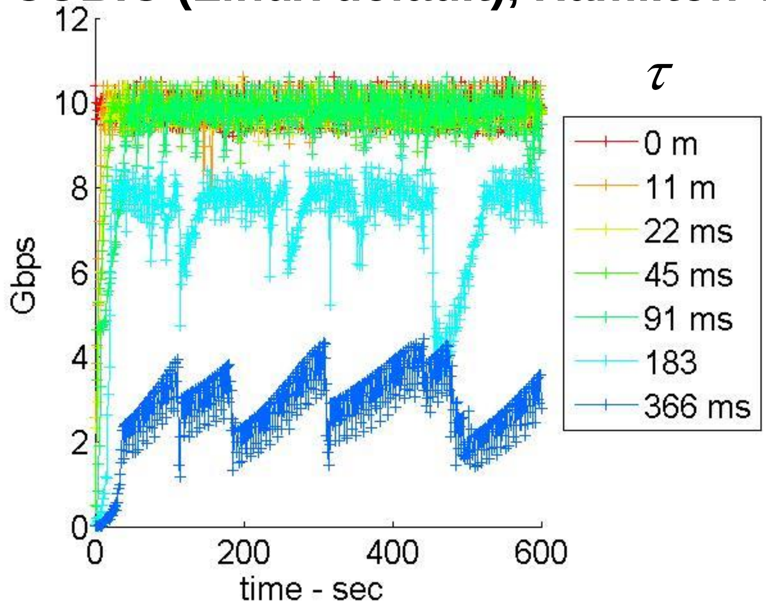


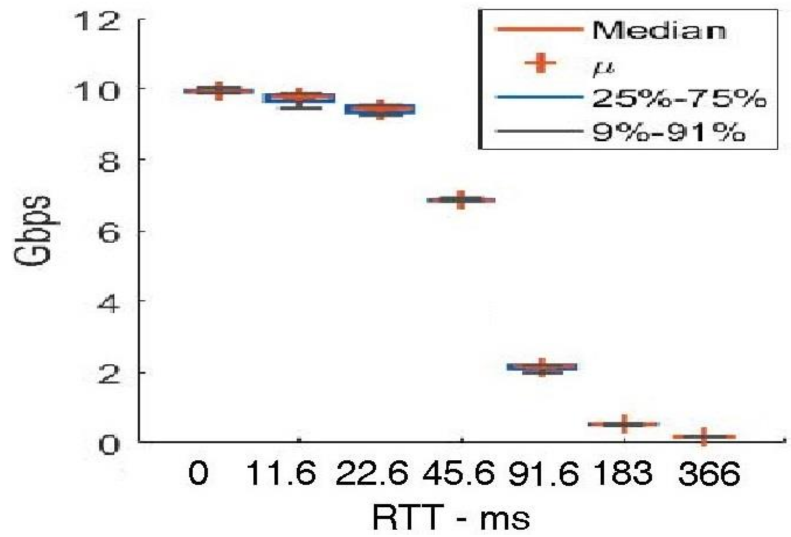
RTTs used in measurements:

- cross-country (0-100ms)
- across continents (100-200ms)
- across globe (366ms)

TCP Memory Throughput Measurements: Uniform Nodes

Throughput traces and profiles: qualitatively similar across TCP variants CUBIC (Linux default), Hamilton TCP, Scalable TCP





Trace:

$\theta(\tau,t)$: **throughput at time** t **over connection with RTT** τ

Throughput Profiles: over period T_O

$$\Theta_O(\tau) = \frac{1}{T_O}\int_0^{T_O}\theta(\tau,t)dt$$

As expected:

- **profile: decreases with RTT**
- **trace: sort of periodic in time**

Not expected:

- **profile: concave at lower RTT**
- **trace: significant variations**
  - **larger at higher RTT**

TCP Throughput Profiles

- **Most common TCP throughput profile**
  - **convex function of rtt**
  - **example, Mathis et al (1997)**

throughput at rtt τ **loss-rate** p

$$\Theta_M(\tau) = \frac{MSS * k}{\tau\sqrt{p}}$$

Function $f(x)$ is *concave* over interval I: for $\tau_1 < \tau_2 \in I$ for all $x \in [0,1]$

$$f\left(x\tau_1 + (1-x)\tau_2\right) \geq xf\left(\tau_1\right) + (1-x)f\left(\tau_2\right)$$

Informally, function is above the linear interpolation

Convex: use $\leq$ in place of $\geq$

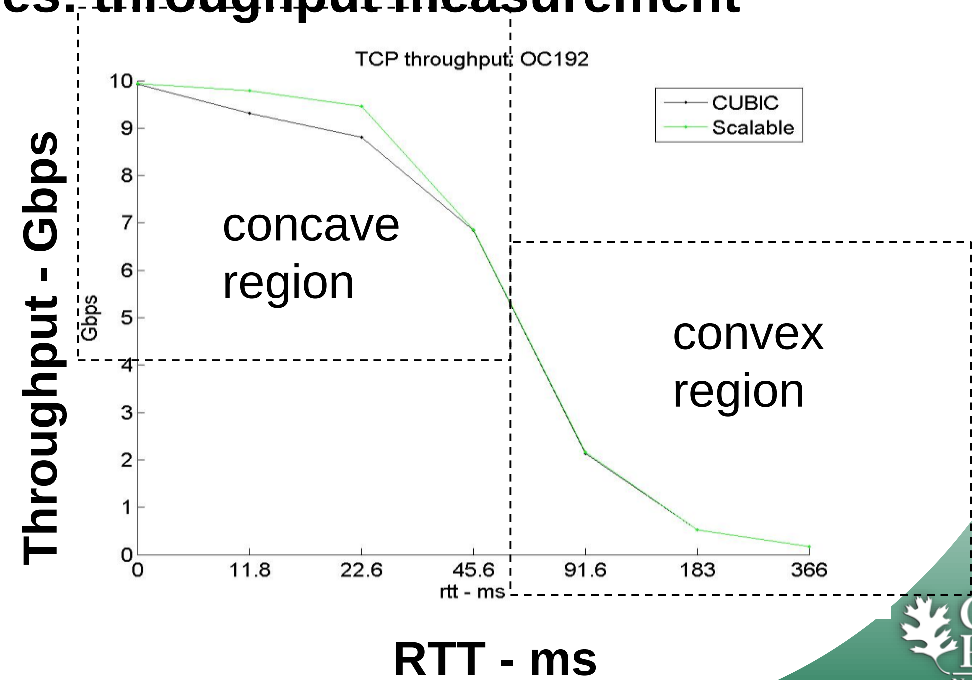
- **Observed Dual-mode profiles: throughput measurement**
  - **CUBIC, STCP, HTCP**

Smaller RTT

- **Concave region**

Larger RTT

- **Convex region**



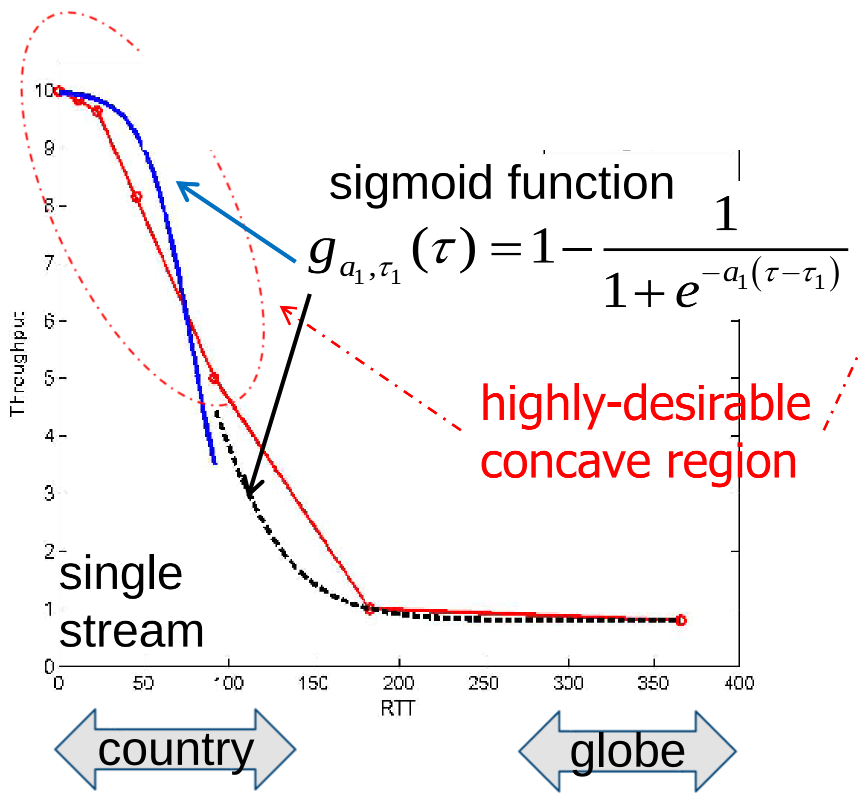
TCP Profiles: memory transfer

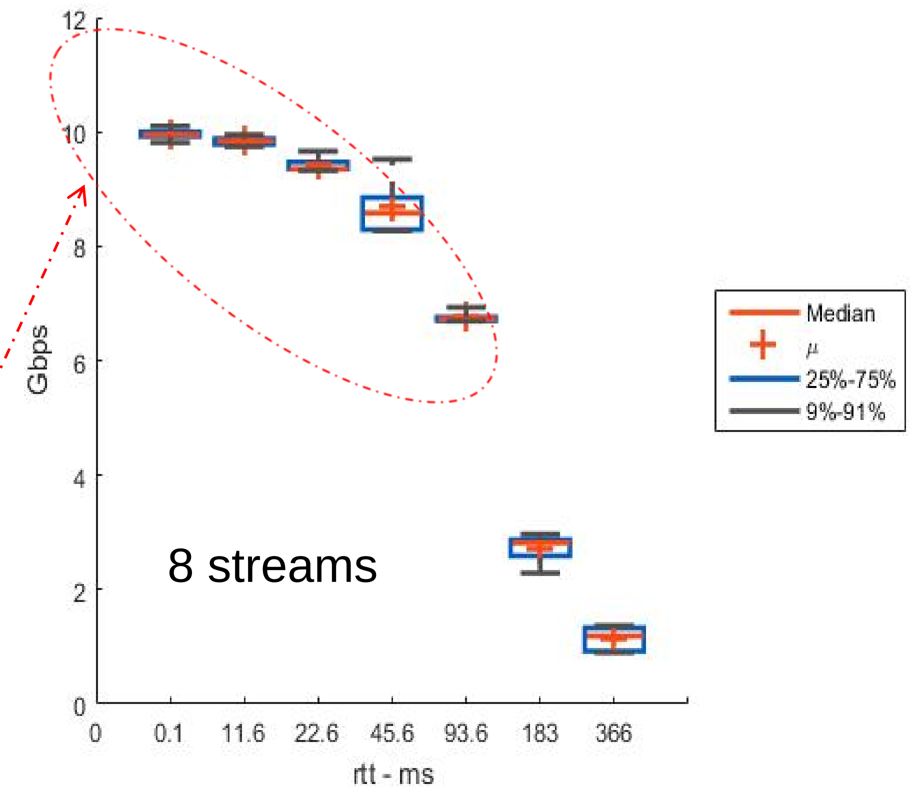
Concave-convex regions – confirmed by sigmoid fits:

10Gbps dedicated connections:

CUBIC congestion control module- default under Linux

- TCP buffers tuned for 200ms rtt: 1-10 parallel streams





RTT: cross-country (0-100ms), cross-continents (100-200ms), across globe(366ms)

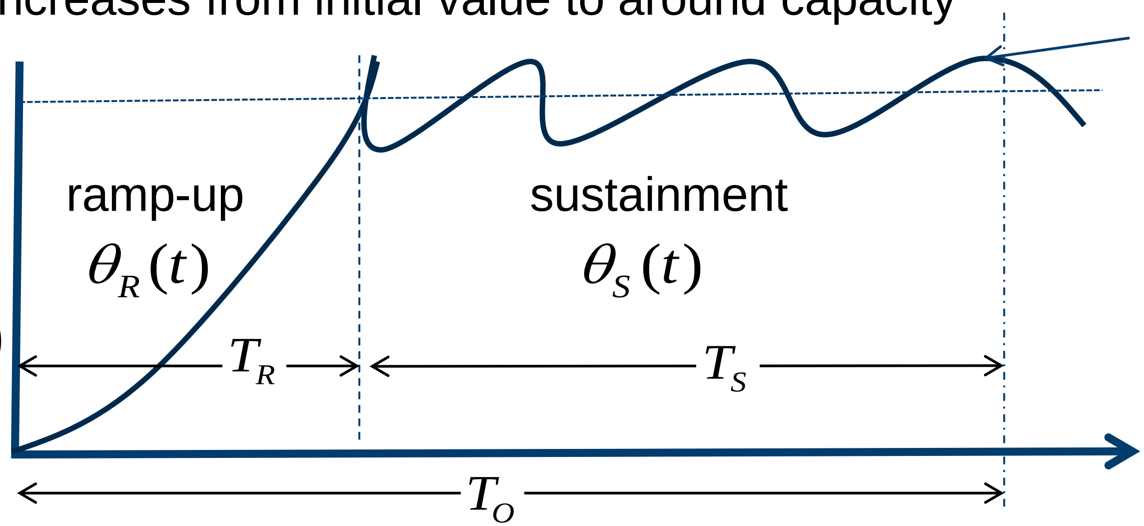
Basic Throughput Model

Throughput trace of *n* streams of TCP version *V* with buffer size *B*: $\theta_V^{B,n}(\tau,t)$

- **Ramp-up Phase**
  - throughput increases from initial value to around capacity

throughput

$$\theta(t)=\theta_V^{B,n}(\tau,t)$$



peak
- connection capacity
- buffer limit

time *t*

$$\bar{\theta}_R(\tau)=\frac{1}{T_R}\int_0^{T_R}\theta(\tau,t)dt \qquad \bar{\theta}_S(\tau)=\frac{1}{T_S}\int_{T_R}^{T_O}\theta(\tau,t)dt$$

- **Sustainment Phase**
  - Throughput is maintained around a peak value $C_\tau^{B,n}$
    - TCP congestion avoidance
  - $\theta_S(t)$ time trace of throughput during sustainment

$$\Theta_O(\tau)=\frac{1}{T_O}\int_0^{T_O}\theta(\tau,t)dt$$

Faster than Slow Start and Multiple TCP flows:

Expand Concavity

Faster than Slow Start:

More increases than slow start: $n_k = \tau^{\in_\tau} \log C \qquad \in_\tau > 0, \tau > 1$

$$T_R = \tau n_k = \tau^{1+\in_\tau} \log C$$

data sent: $1 + 2 + \cdots + 2^{n_k} = 2^{n_k+1} - 1 = 2^{1+\tau^{\in_\tau}} C - 1$

$$\bar{\theta}_R \approx \frac{2^{1+\tau^{n_k}} C}{\tau^{1+\in_\tau} \log C}$$

Average Throughput:

$$\Theta_O(\tau) = \frac{2^{1+\tau^{\in_\tau}} C}{T_O} + C\left[\frac{T_O - \tau^{1+\in_\tau} \log C}{T_O}\right]$$

$$\frac{d\Theta_O}{d\tau} = -\underbrace{\frac{(1+\in_\tau)\tau^{\in_\tau} C \log C}{T_O}}$$

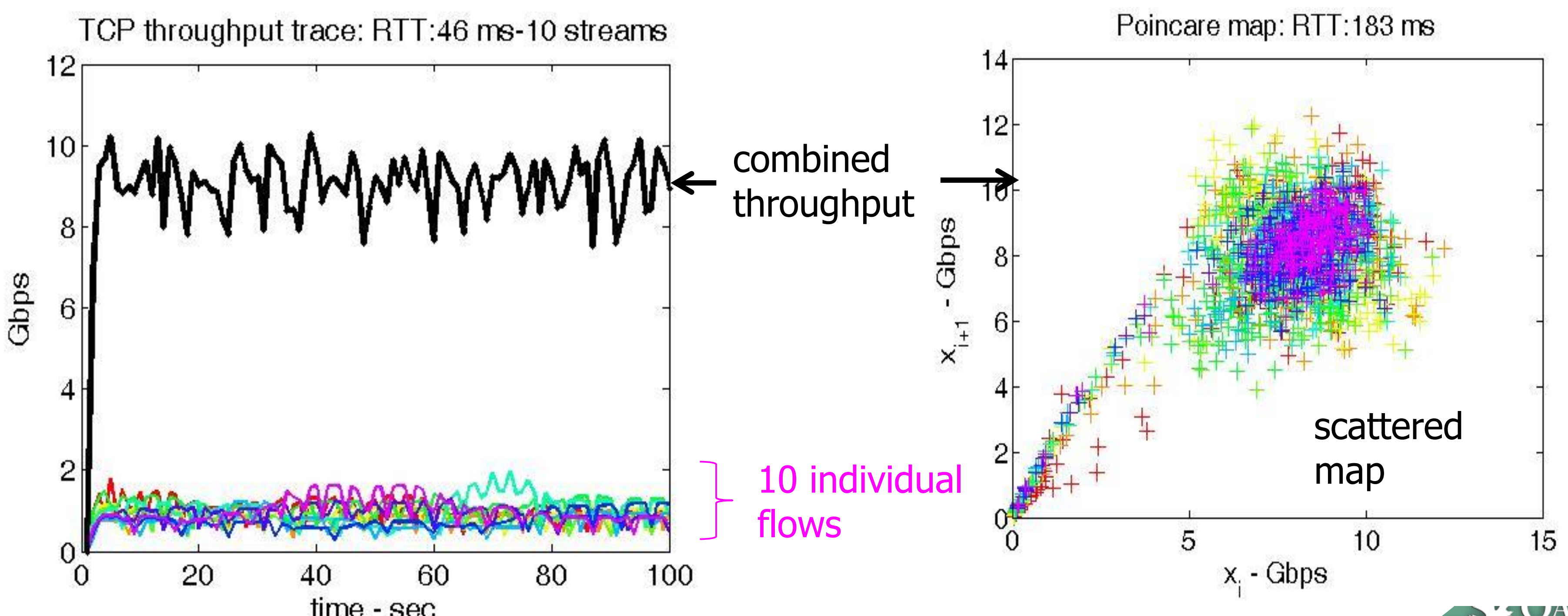
decreasing function of τ

implies concavity of $\Theta_O(\tau)$

Poincare Map

Well-Known tool for analyzing time series – used in chaos theory

- **Poincare map** $M:\Re^d \rightarrow \Re^d$
  - **Time series:** $X_0, X_1, \cdots, X_i, X_{i+1}, \cdots$
  - **generated as** $X_{i+1} = M\left(X_i\right)$
- **Effect of Poincare map:**
  - **range specifies achievable throughput**
  - **complexity indicates rich dynamics – lower throughput and narrow concave**



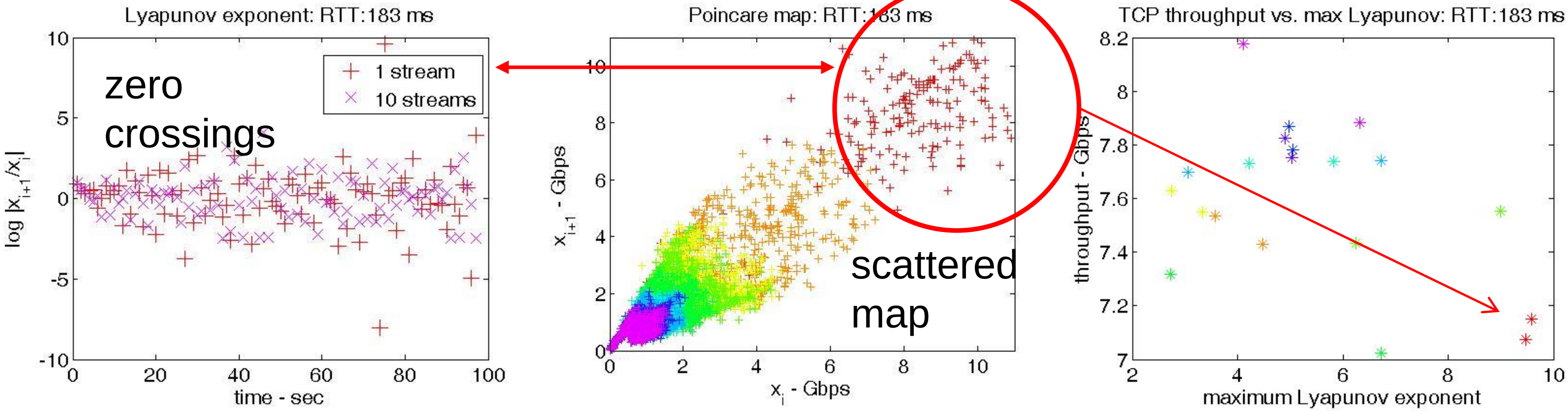
Lyapunov Exponent: Stability and Concavity

- **Log derivative of Poincare map**

$$L_M = \ln\left|\frac{dM}{dX}\right|$$

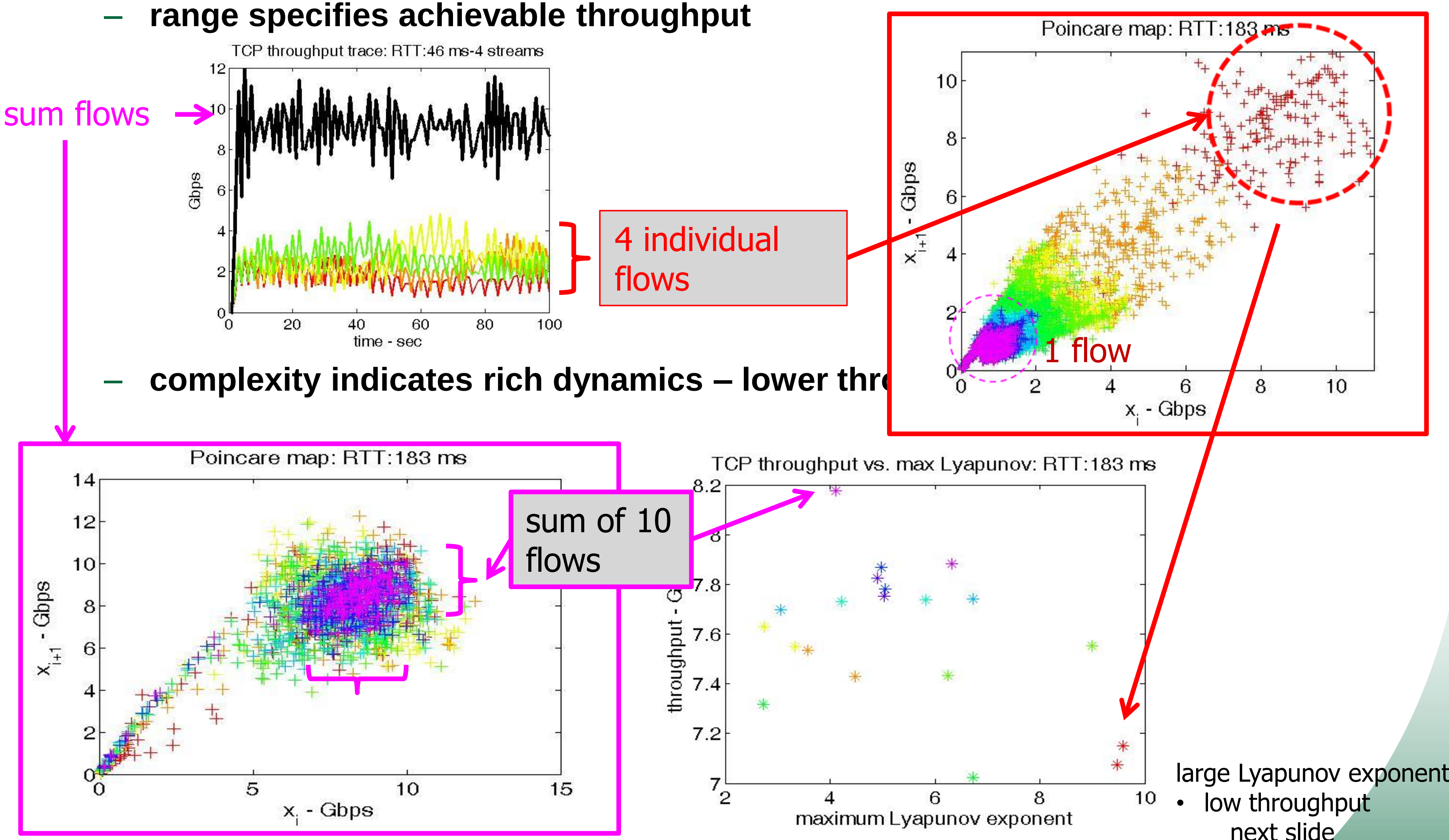
- **Provides critical insights into dynamics**
  - **Stable trajectories:** $L_M < 0$
  - **Chaotic trajectories:** $L_M > 0$
    - **indicate exponentially diverging trajectories with small state variations**
    - **larger exponents indicate large deviations**
  - **protocols are operating at peak at rtt**
    - **stability implies average close to peak - implies concavity**
    - **positive exponents imply lowered throughput – trajectories can only go down**
      - » **then, weak sustainment implies convexity**

large L_M implies low throughput



Poincare Map and Lyapunov Exponent

- **Effect of Poincare map:**
  - **range specifies achievable throughput**
  - **complexity indicates rich dynamics – lower thr**



Instability shrinks concave region

Two protocols P_1 and P_2 with Lyapunov exponents L_1 and L_2

Consider $L_1 > L_2$

Trajectories of P_1 deviate faster than those of P_2 both operating at peak

which implies $\bar{\theta}_S^1 \leq \bar{\theta}_S^2$

For fixed $\bar{\theta}_S$

we have $$\frac{\partial \Theta_O}{\partial \tau} = -\frac{\partial f_R}{\partial \tau}\left(\bar{\theta}_S - \bar{\theta}_R\right)$$

since $\frac{\partial f_R}{\partial \tau} \geq 0$, concavity of Θ_O is equivalent to condition $\left(\bar{\theta}_S - \bar{\theta}_R\right) > 0$

for a fixed configuration, the condition $\bar{\theta}_S^1 \leq \bar{\theta}_S^2$ leads to

$$\left\{\tau : \bar{\theta}_S^1 \geq \bar{\theta}_R\right\} \subseteq \left\{\tau : \bar{\theta}_S^2 \geq \bar{\theta}_R\right\}$$

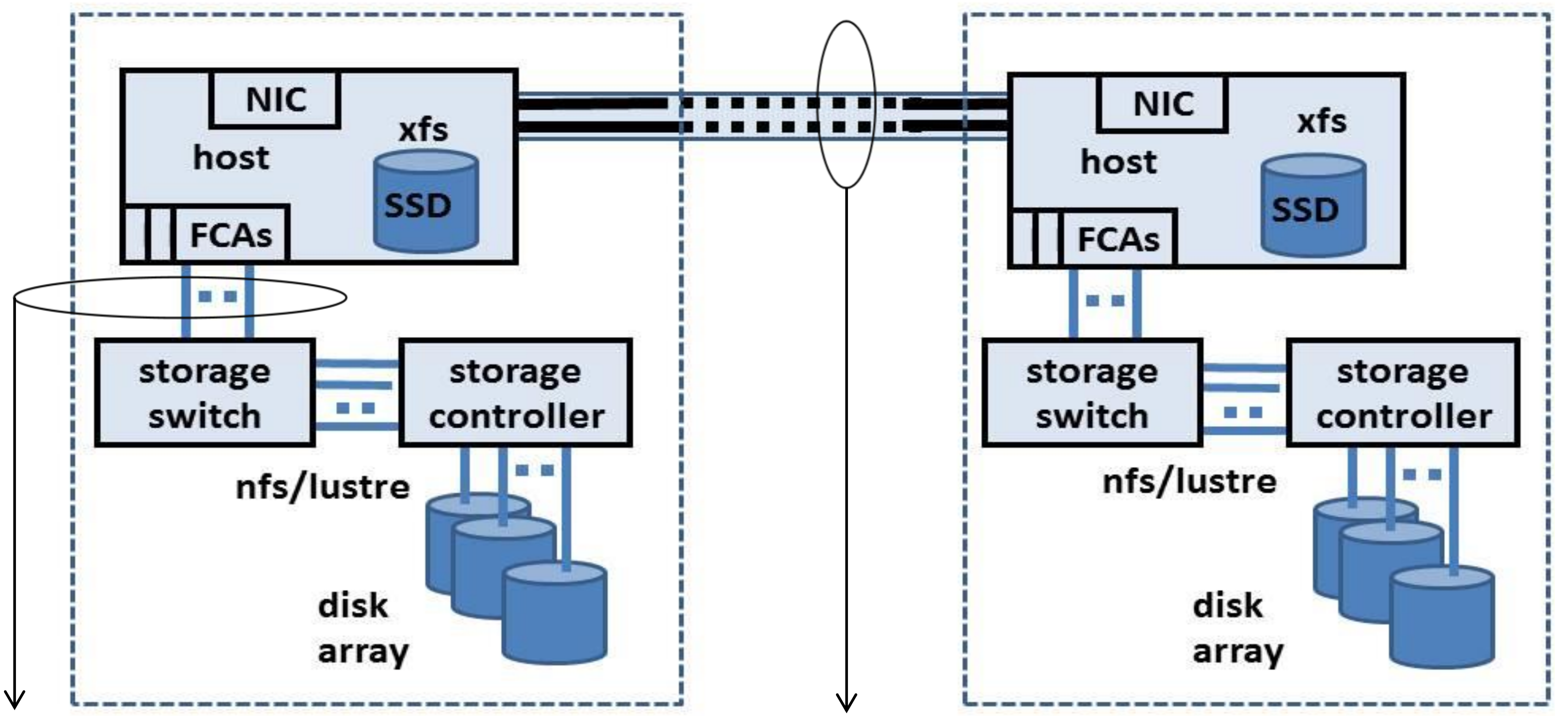
which implies P_2 has larger concave region

In general, stable throughput dynamics are highly desirable for achieving

(a) peak throughput, and (b) concave throughput profiles

Informally, both start at around peak, P_1 becomes lower faster
- switches to convex compressing the concave region

Network and IO Systems:
Wide-area file transfers involve complex systems



Peak IO rates: xddprof on hosts
- xfs: ~40 Gbps
- lustre: ~32 Gbps

Peak n/w throughput:
- iperf: 0ms rtt
  - TCP: > 9Gbps
  - UDP/T: > 8Gbps

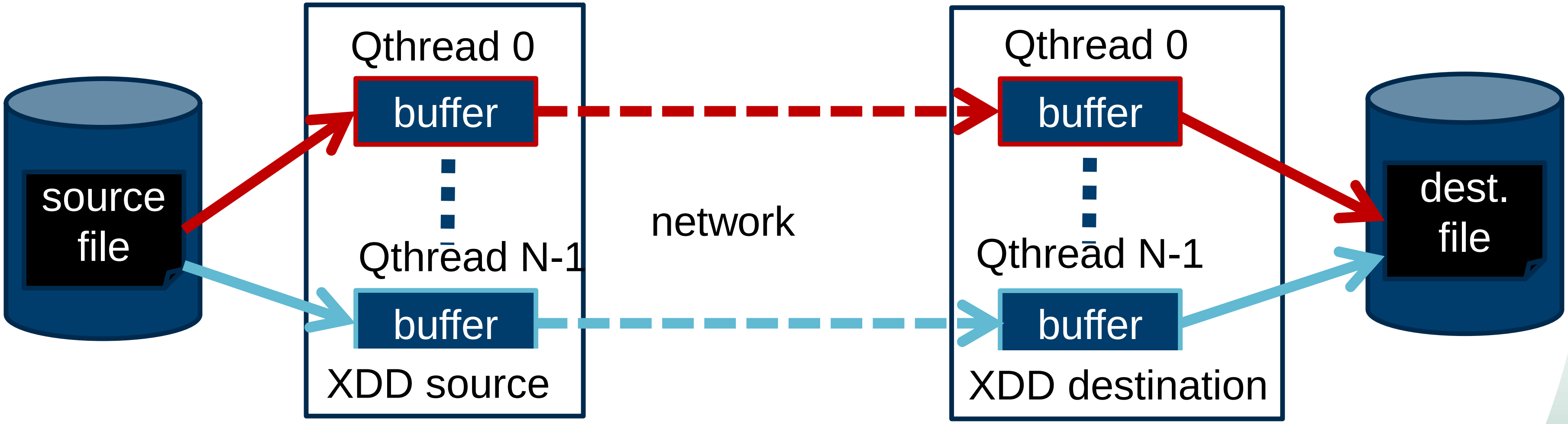
Individually, more IO flows lead to higher IO throughput
more parallel TCP streams lead to higher network throughput
But, not necessarily true when composed: 8 Lustre IO + 8 TCP flows : 2Gbps

XDD: host-to-host file transfer tool

- XDD uses parallel flows to move files
  - each flow is composed of
    - source IO/file flow + TCP flow + destination IO/file flow
    - data is read/written in blocks – sizes 8k,65k, 148k

Intuitively, more flows must provide high file transfer rate

But, our measurements show more complex dependency



Lustre file parameters
- stripe number: #OST
- stripe size

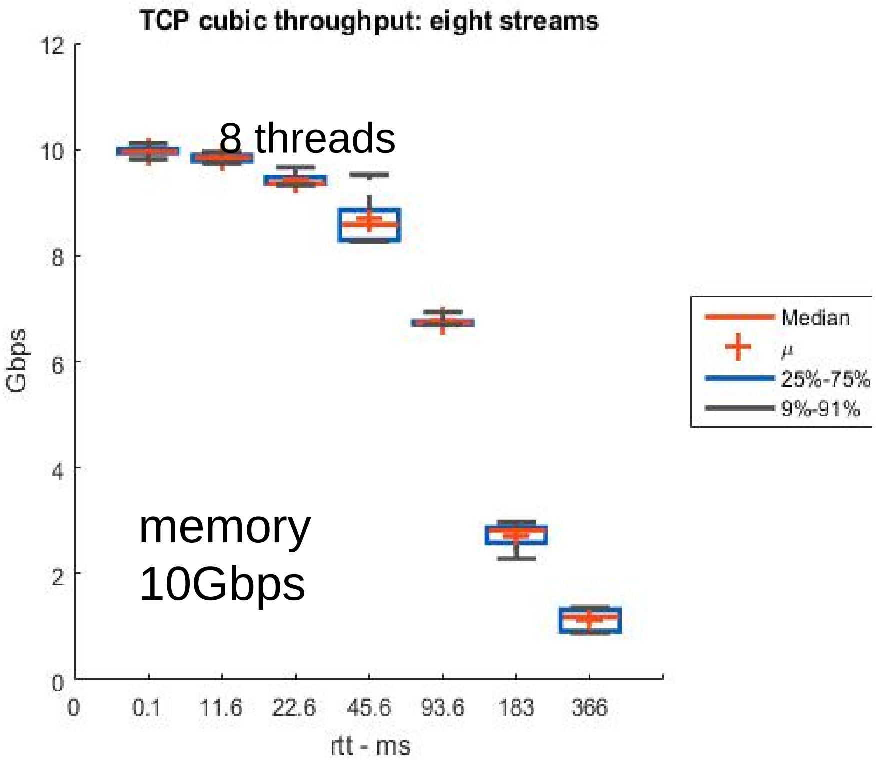
XDD:

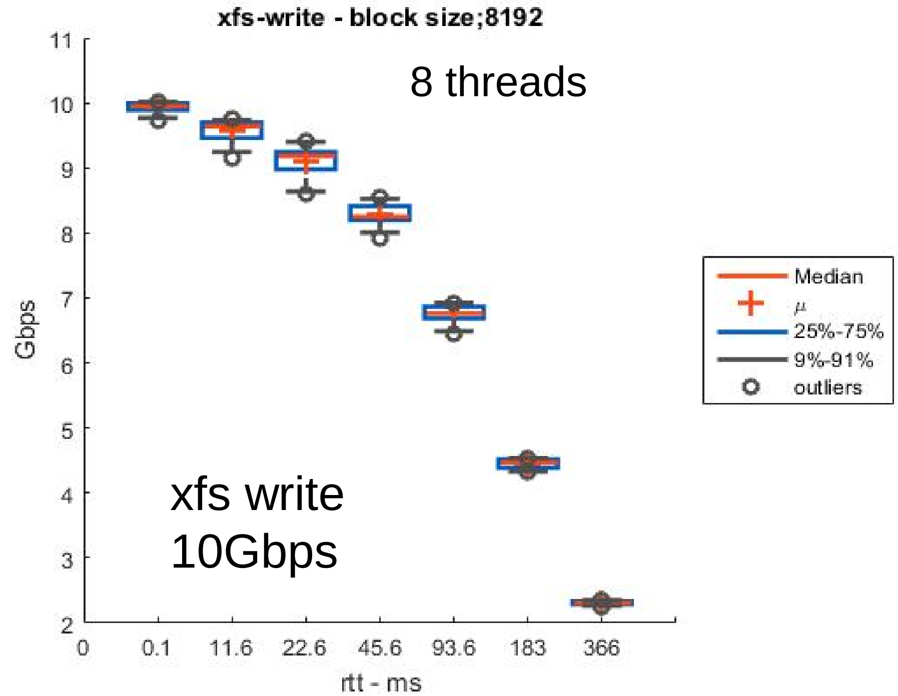
#IO threads = # parallel TCP streams = #flows

Utilizes its own buffers – different from Lustre buffers

TCP CUBIC and xfs file systems

- **xdd host-to-host file transfers: peak: 10Gbps**





xdd file IO throughput is close to TCP throughput

- 8 IO threads and 8 TCP parallel streams
- Impedance mismatch is quite small

Lustre Over Wide-Area

Lustre distributed file system

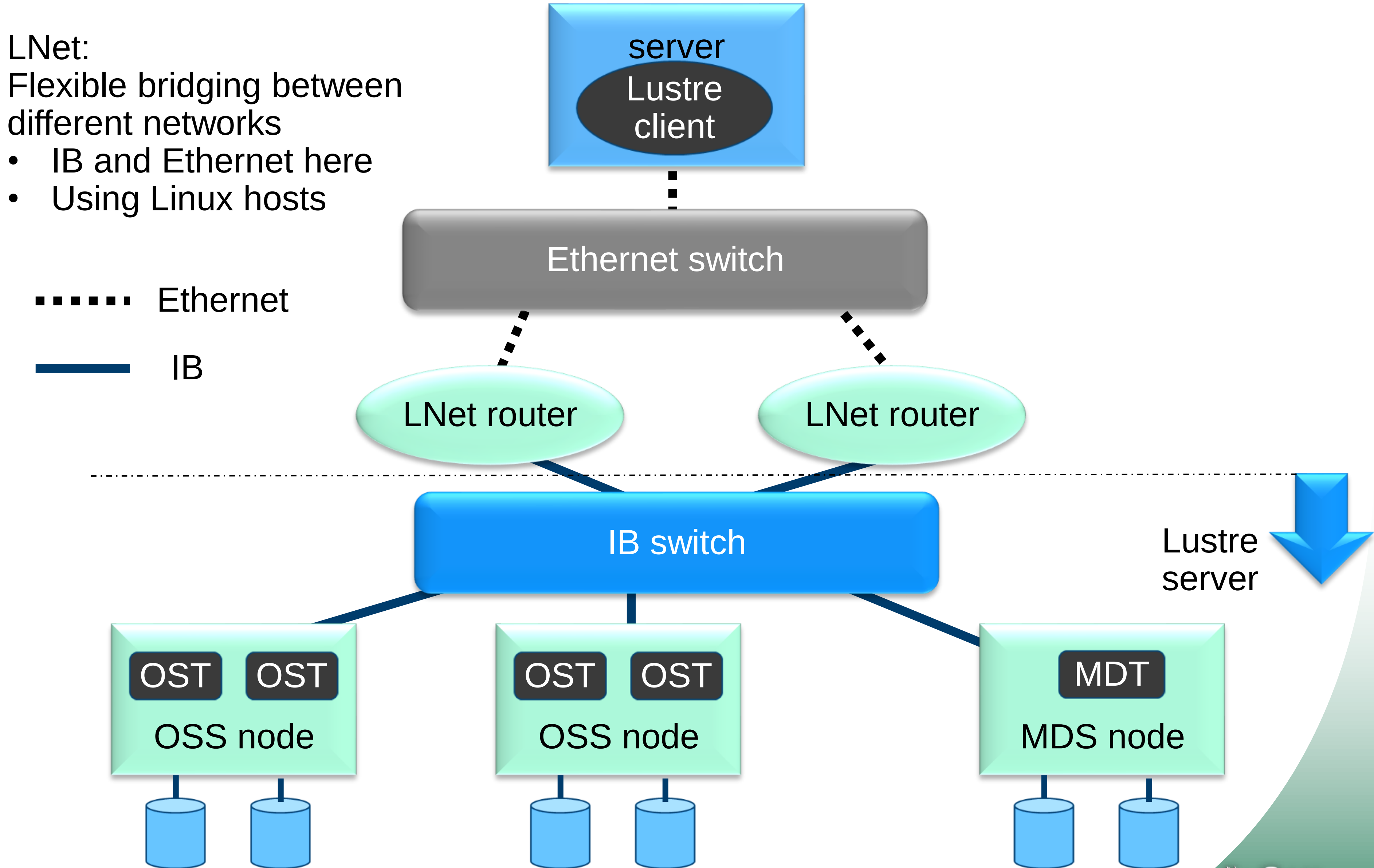
- Meta Data Servers (MDS)
- Object Storage Servers (OSS)
  - supported by one or more Object Storage Target (OST)
- High performance: parallelizing I/O from multiple clients to multiple OSTs: striped files

- **Desired: Lustre mounted over wide-area**
  - No need for transfer services such as GridFTP, Aspera, XDD and others
  - Easier application integration with remote file operations
- **Current Installations**
  - Majority: over site IB networks: Time-out limitation: 2.5ms
  - IB WAN extenders: too expensive and not flexible
- **Solution: Lustre over Ethernet (not as widely deployed)**
  - TCP/IP implementation: uses existing networks
  - Very little infrastructure enhancements needed

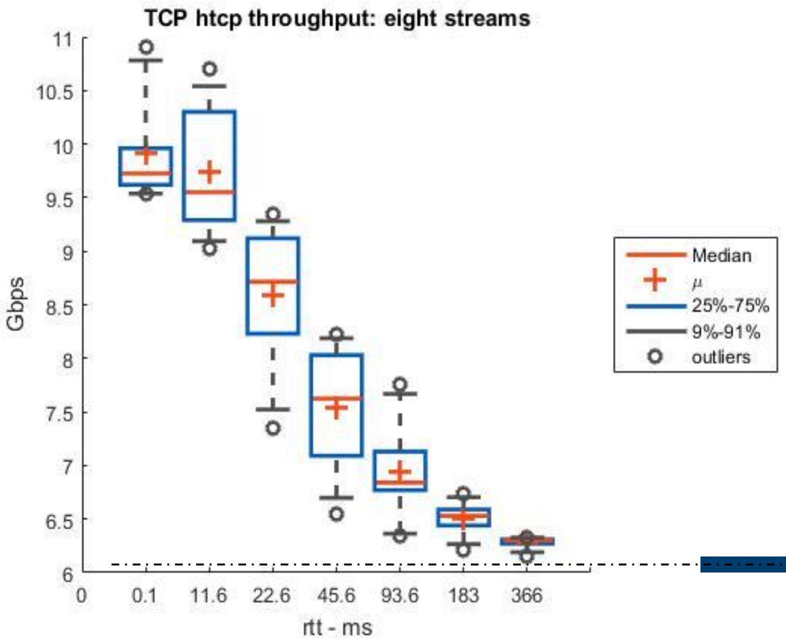
Lustre over IB-Ethernet: LNet routers

LNet:
Flexible bridging between different networks

- IB and Ethernet here
- Using Linux hosts

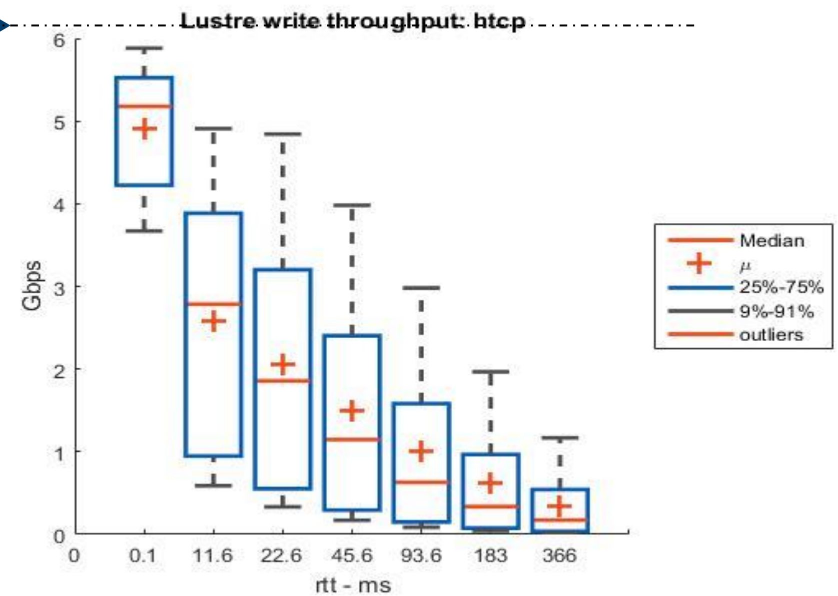


Lustre wide-area: bohr – Hamilton TCP



bohr IO servers – TCP tuned

- 48core, opteron, 2.2 GHz
- write peak throughput: ~6Gbps
- lower than lowest iperf throughput
- Centos 6.8



Hamilton TCP: Not much difference

Recommended for large transfers over long (cross-country and inter-continental) distances

- Used in DOE Data Transfer Nodes (DTNs)
- CUBIC is default in Linux

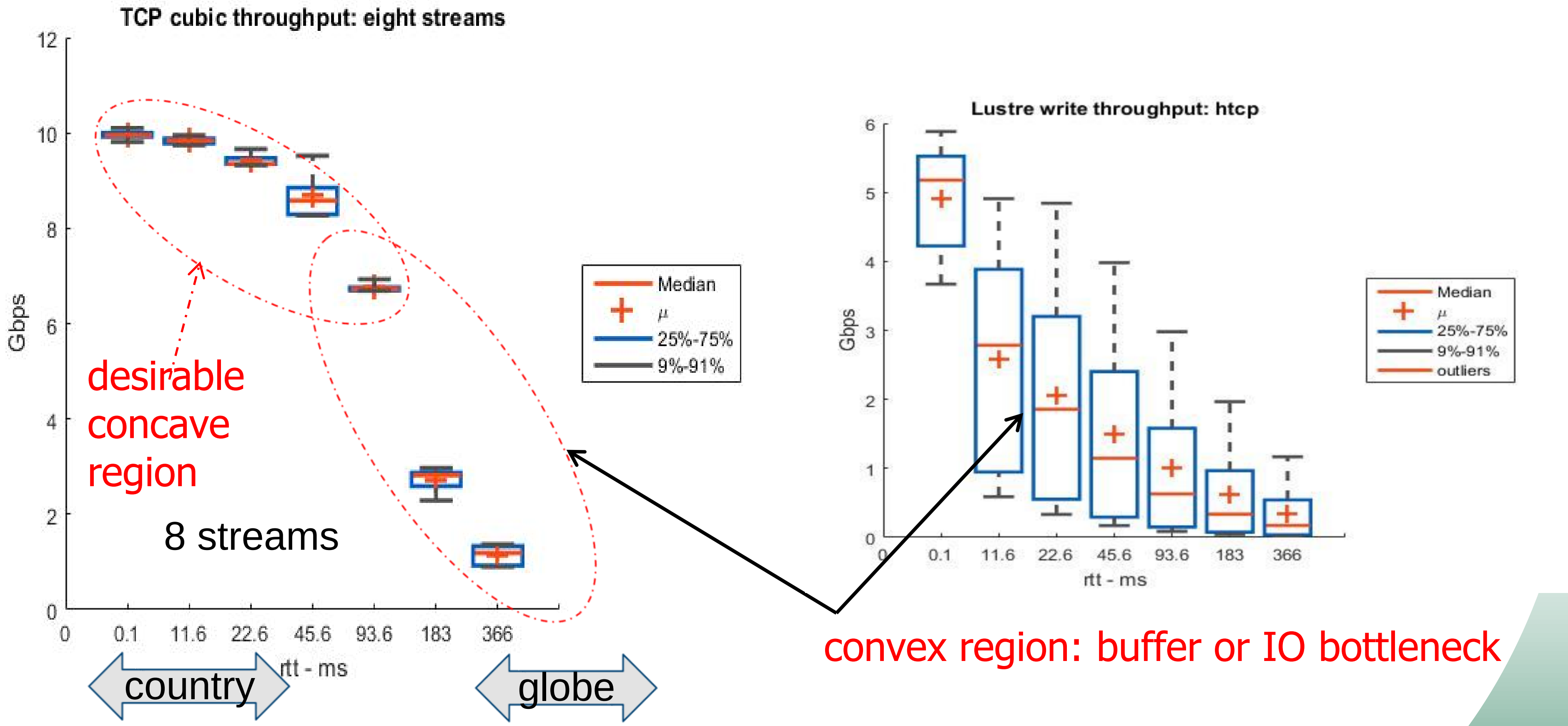
IO or Network Bottleneck?

TCP memory transfers: concave-convex regions

10Gbps: CUBIC TCP buffers tuned for 200ms rtt

Concave region: indicates buffer, IO bottleneck

Our Lustre configuration indicates IO limit



RTT: cross-country (0-100ms), cross-continents (100-200ms), across globe(366ms)

Generic Model for Data, Disk and File Transfers

Buffer size, IO throughput or available processing power limit data in transit:

connection capacity (bps): C

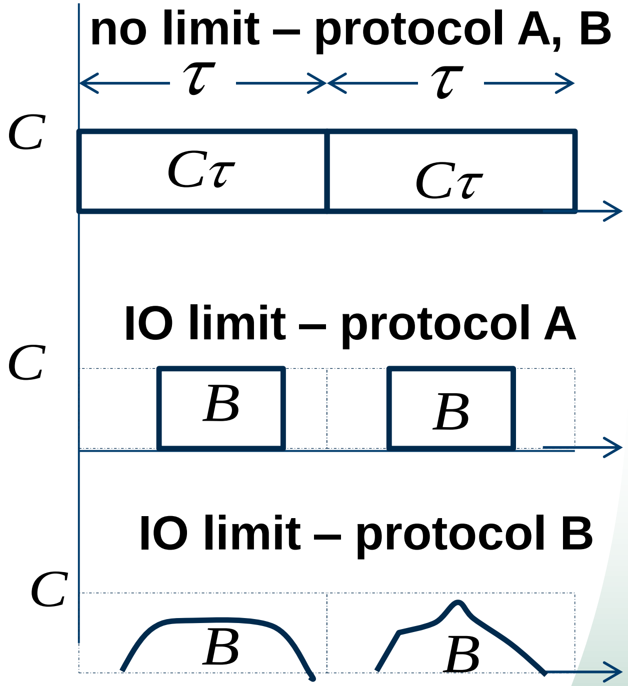
RTT: τ

data unacknowledged within a slot of period: τ

no IO or processor limit: $C\tau$

under IO or processor limit: $B < C\tau$

example: limited buffer size



Throughput averaged over each slot of width τ:

$$\theta^{\tau}(t) = \frac{B}{\tau}$$

Throughput profile: $$\Theta_O(\tau) = \frac{1}{T_O}\int_0^{T_O} \theta^{\tau}(t)dt = \frac{B}{\tau}$$

Throughput derivative: $$\frac{d\Theta_O}{d\tau} = -\frac{B}{\tau^2}$$

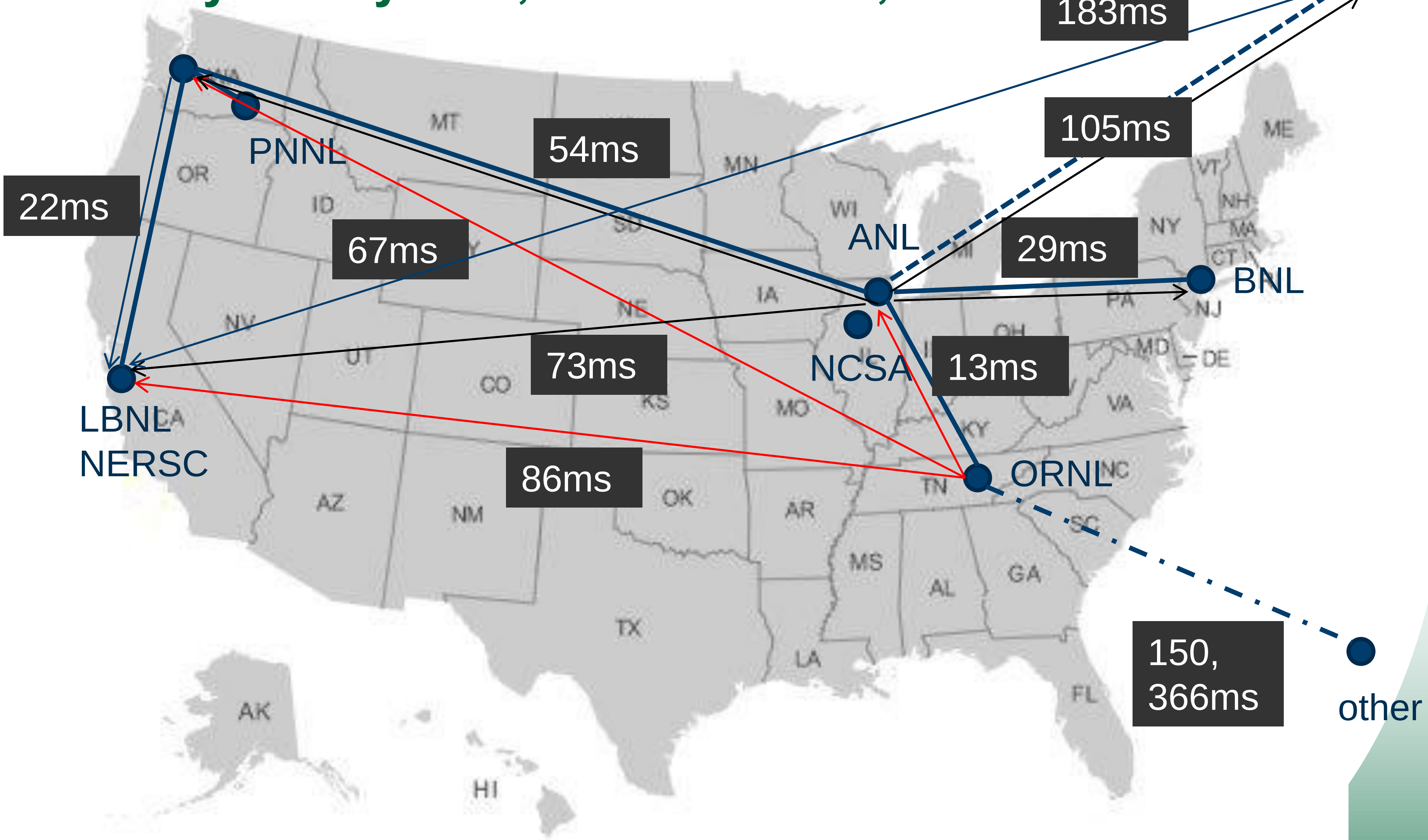
increasing function of τ **implies convexity of** $\Theta_O(\tau)$

Transport methods may have different shapes of B – but subject to convexity

- **convex profile indicates disk or file throughput limit**
- **due to peer credits on IB and Ethernet sides of LNet**

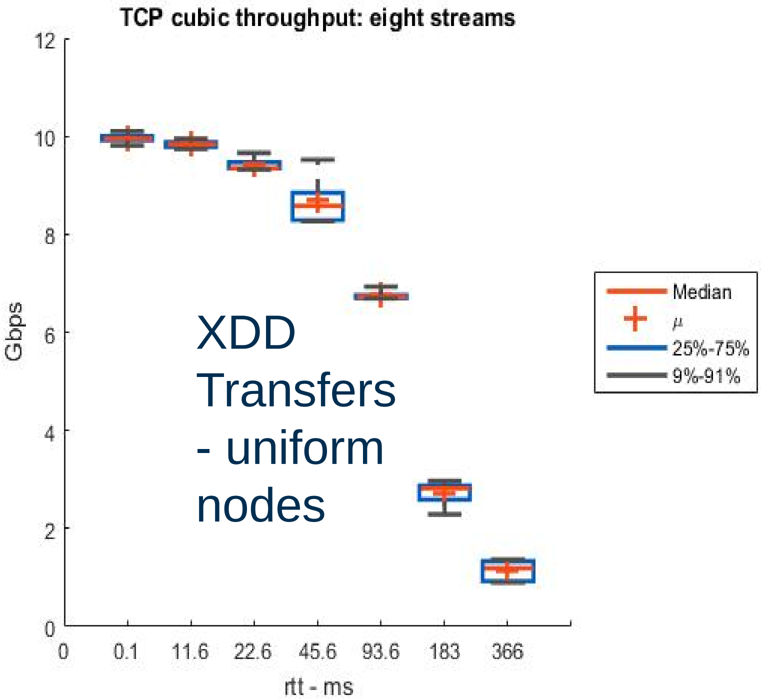
Data transfer infrastructures:

Sites vary: file system, transfer hosts, …

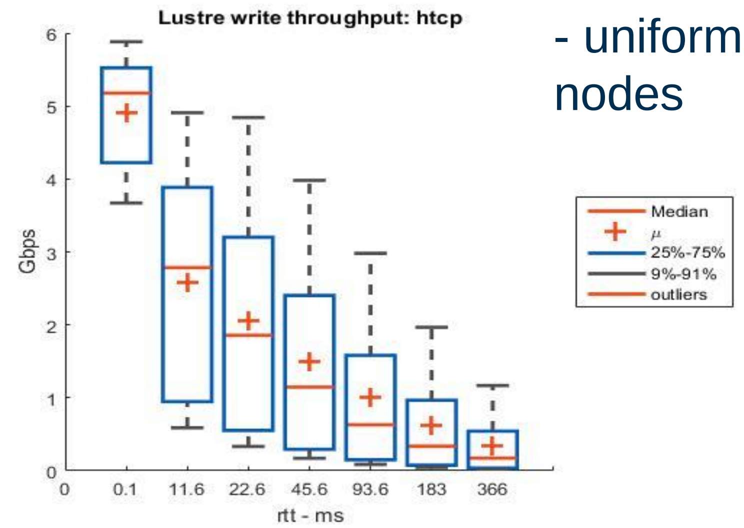


Profiles of infrastructures

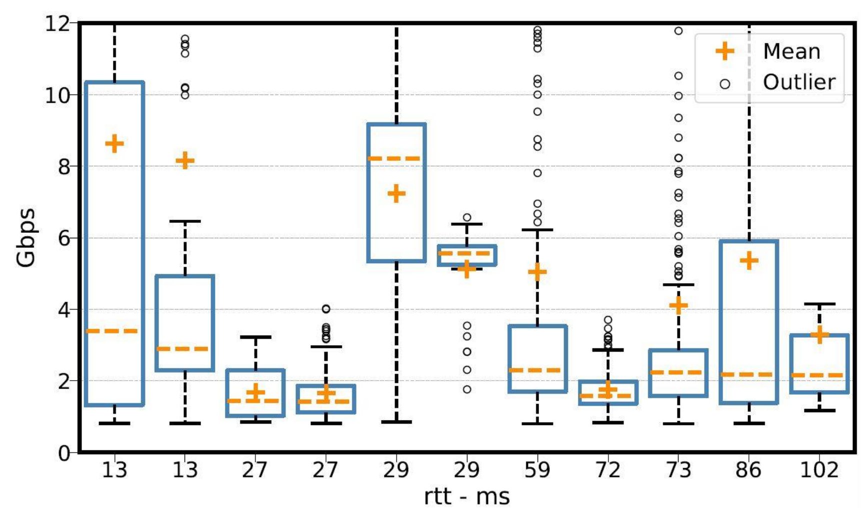
XDD Transfers
- uniform nodes



LNet Lustre
- uniform nodes

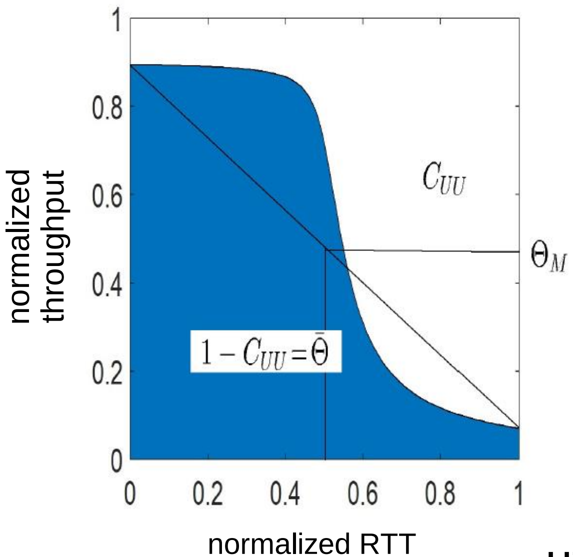


Globus file transfers
- production infrastructure
- site variations lead to complex profiles



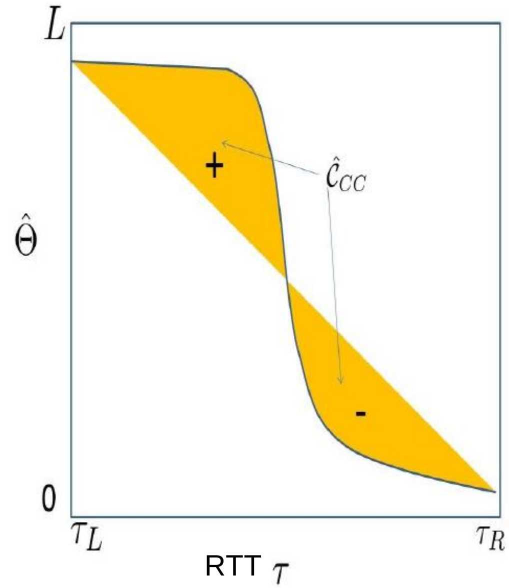
Utilization-Concavity Coefficient

- **Scalar** $C_{UC} \in [0, 1]$
  - **Normalized with respect to throughput and rtt**
  - **Incorporates both concavity and utilization throughput profiles**



$$\mathcal{C}_{CC}\left(\tilde{\Theta}\right) = \bar{\tilde{\Theta}} - \tilde{\Theta}_M$$

$$\mathcal{C}_{UC}\left(\tilde{\Theta}\right) = \bar{\tilde{\Theta}} - \tilde{\Theta}_M/2 + 1/4$$

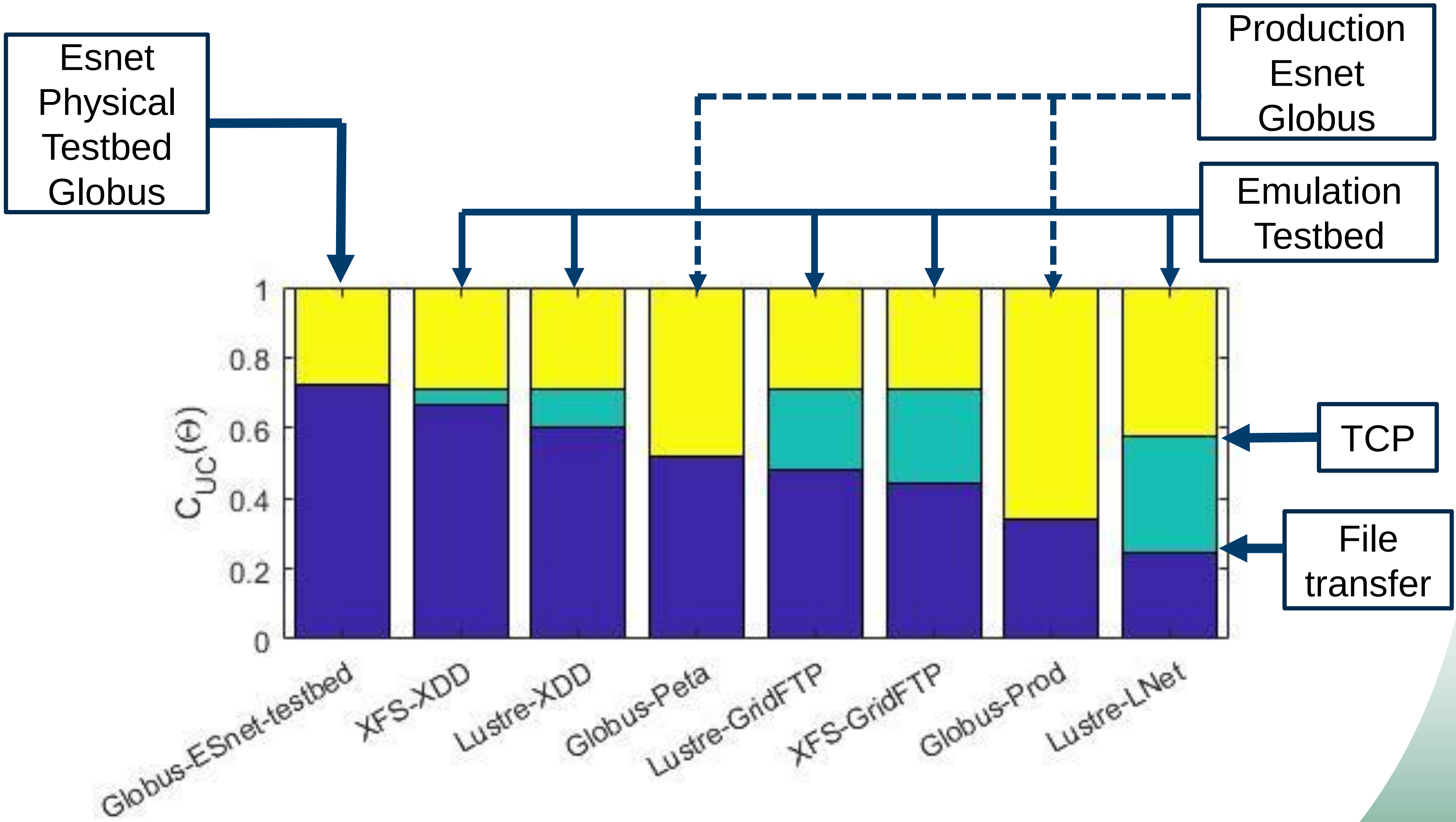


$$\mathcal{C}_{UC}(\Theta) = \frac{1}{2}\left(\underbrace{[1 - \mathcal{C}_{UU}(\Theta)]}_{\text{utilization}} + \underbrace{\left[\frac{1}{2} + \mathcal{C}_{CC}(\Theta)\right]}_{\text{concavity}}\right)$$

utilization

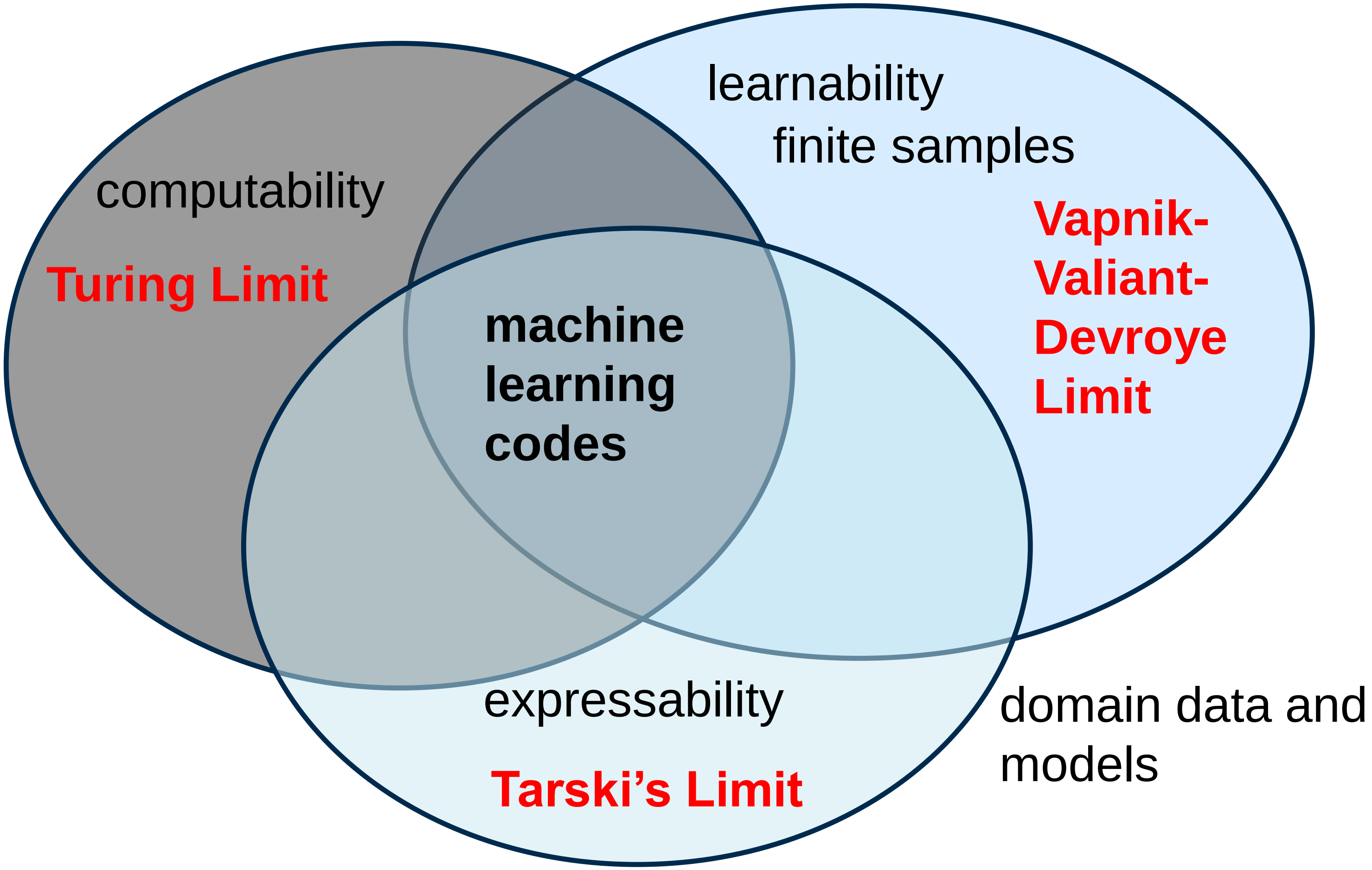
concavity

Coefficient for 8 different transport infrastructures



Foundational Limits of Machine Learning Codes

Computations executed on machine with data and models



Throughput profiles have monotonicity properties: effectively learnable

Confidence Estimates

$\theta(\tau,t)$: random with distribution $P_{\Theta_O(\tau)}$ that depends on

- TCP version and parameters
- host and connection parameters

Profile regression: $\bar{\Theta}_O(\tau) = E\left[\Theta_O(\tau)\right] = \int \Theta_O(\tau) P_{\Theta_O(\tau)}$

Profile mean based on measurements: $\theta(\tau_k, t_i^k) : k = 1, 2, \cdots, n; i = 1, 2, \cdots, n_k$

$$\hat{\Theta}_O(\tau_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \theta(\tau_k, t_i^k)$$ ← machine learned profile

Estimate of profile regression f chosen from class of monotone functions M

TCP profile decreases with RTT

Error of estimate $I(f) = \int \left[f(\tau) - \theta(\tau,t) \right]^2 P_{\theta(\tau,t)}$

Best estimate: $f^* : I(f^*) = \min_{f \in M} I(f)$

Linear interpolation based on profile mean is close to optimal probabilistically

$$P\left\{ I(\hat{\Theta}_O) - I(f^*) > \in \right\} < \delta \qquad \delta = 32 \left(\frac{n}{\in} \right)^{(1+C/\in)\log_2(4\in/C)} n e^{-\in^2 n/(2C)^\wedge 2}$$

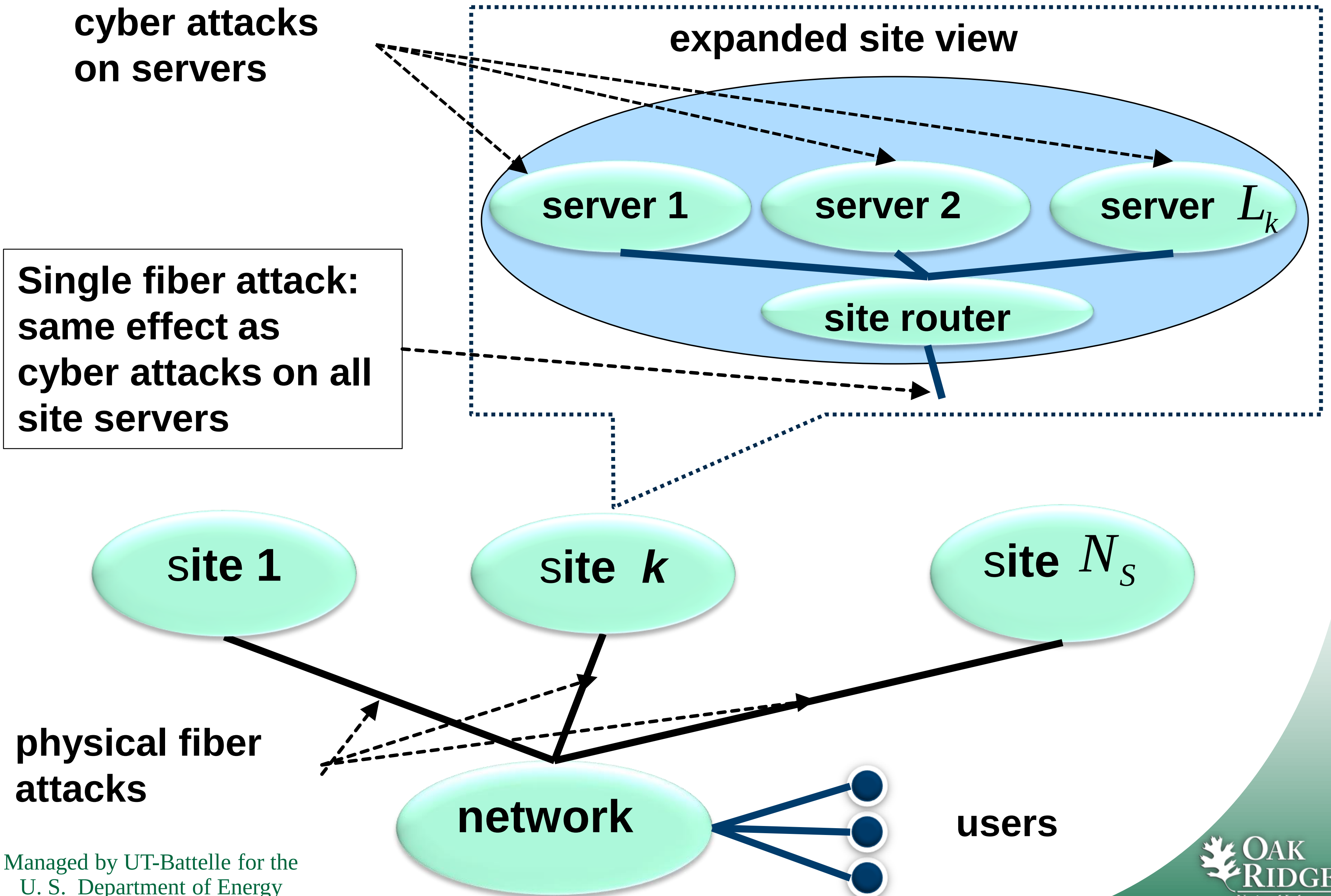
Gets better with more measurements

Intuitively, profile is close to optimal with high probability

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

OAK RIDGE National Laboratory

Multi-Site Cloud Computing Infrastructure



Infrastructure: Systems of Components

Consists of N individual systems: $S_1, S_2, \ldots, S_N$

each system consists of cyber and physical components

x_i **:defenders investment in system in defending** S_i

example: number of reinforced components of S_i

y_i **:attackers investment in attacking system** S_i

example: number of reinforced components of S_i

P_i **:survival probability of system** S_i

example: contest success function $P_i = \frac{x_i^m}{x_i^m + y_i^m}$

P_I **:survival probability of multiple system infrastructure**

In general, it depends on:

defenses $x_1, x_2, \ldots, x_N$

attacks $y_1, y_2, \ldots, y_N$

correlations

This formulation captures cloud computing infrastructure

- Not flexible to capture varying complexities of systems

Defender Utility: General Form

Defender minimization utility function:

$$
\begin{aligned}
U_D\left(x_1,\ldots,x_{N_S},y_1,\ldots y_{N_S}\right) \\
&= F_{D,G}\left(x_1,\ldots,x_{N_S},y_1,\ldots y_{N_S}\right)G_D\left(x_1,\ldots,x_{N_S},y_1,\ldots y_{N_S}\right) \quad \text{reward term} \\
&+ F_{D,L}\left(x_1,\ldots,x_{N_S},y_1,\ldots y_{N_S}\right)L_D\left(x_1,\ldots,x_{N_S}\right) \quad \text{cost term}
\end{aligned}
$$

Defender: reinforces

x_i **number of components reinforced of basic system** S_i

Attacker:

y_i **number of components attacked of basic system** S_i

Infrastructure Survival Probability Estimate at Nash Equilibrium:

Defender's estimates of survival probability of system $S_b; b = 1, 2, \ldots, N$

$$\hat{P}_{b;D} = \frac{\dfrac{\partial C_b}{\partial x_b} + \dfrac{F_{G,L}^{D,b}}{L_{G,L}^{D}}}{\dfrac{\partial C_b}{\partial x_b} - (1 - C_b)\Lambda_b}$$

Simple dependence on:

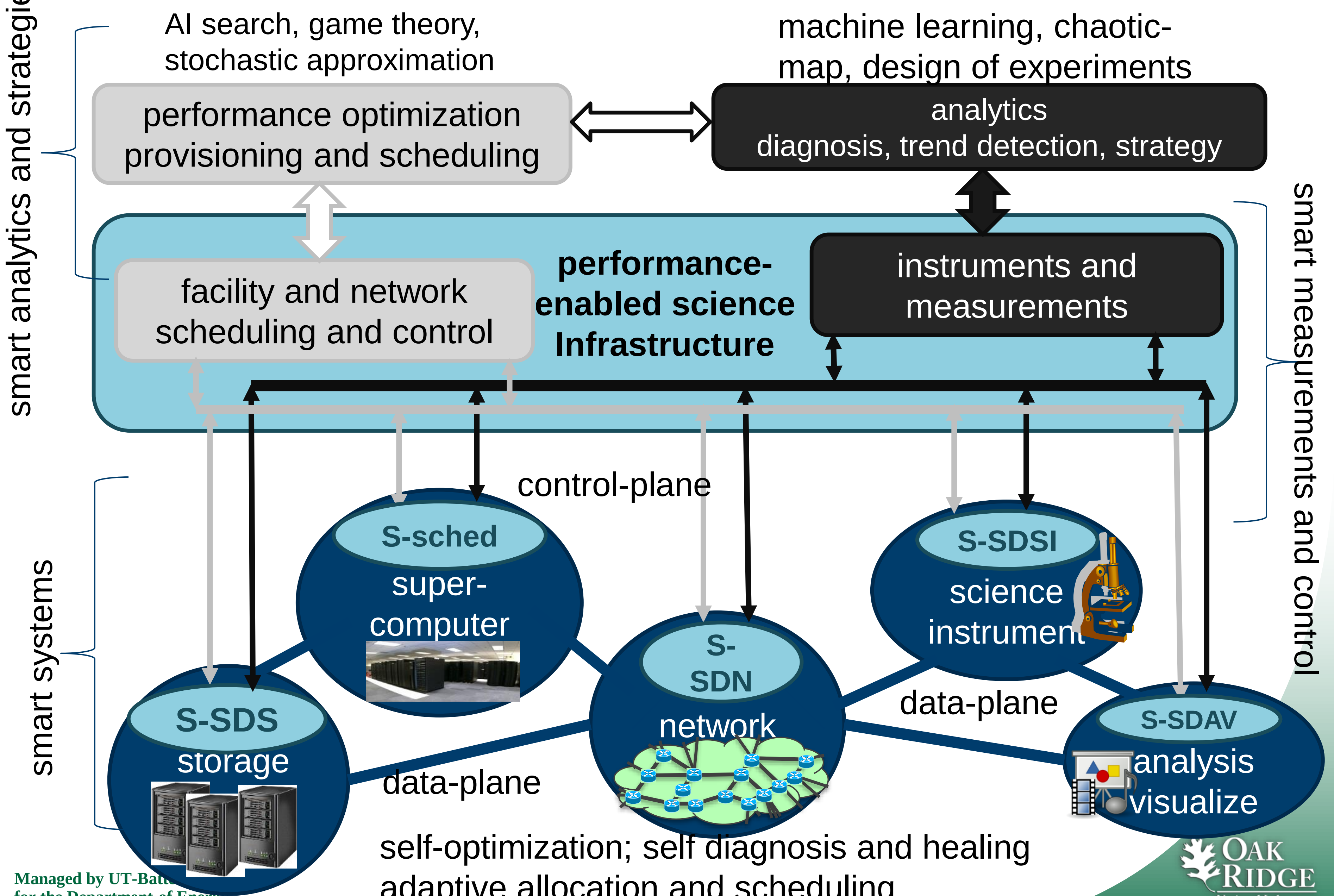
- correlation function: between systems
- multiplier function components within systems

under the condition: $C_b < 1$ **or** $\dfrac{\partial C_D}{\partial x_b} \neq 0$

Observations: Survival probability estimates depend

- **gain-cost and gain-cost gradient**
- **aggregate correlation function and its derivative**
- **system multiplier functions**

Looking into Future:
Integrated Softwarized Federated Instruments for Science:



Looking into Future

Scientific Methods: important in design, analysis and optimization of data transport across time-space distance:

- Profile estimation and performance optimization
- Analytics, machine learning, measurements design, game theory, …

Industry is developing powerful solutions

- Softwarization, virtualization, containerization, …
- Transport tools, methods, TCP versions, UDP transport, …

But, their main targets are:

- Cloud computations with large number of users – optical networks
- IOT with larger number of devices – wireless networks

But, special infrastructures are outside industry's path

- Small number of large sources over optical networks
- Data transport, streaming, computational monitoring and steering, interactive remote experiments

Focused support needed

- will not happen as industrial by-products
- efforts similar to HPC systems needed to foster this area

New science of data over time-space distance: components and infrastructures

Thank you

References

1. N. S. V. Rao, C. Y. T. Ma, K. Hausken, F. He, D. K. Y. Yau, J. Zhuang, Defense strategies for asymmetric networked systems with discrete components, Sensors, vol. 18, 2018, pp. 1421.
2. N. S. V. Rao, SDN solutions for switching dedicated long-haul connections: Measurements and comparative analysis, International Journal on Advances in Networks and Services, vol. 9, no. 3-4, 2016.
3. N. S. V. Rao, Q. Liu, S. Sen, R. Kettimuthu, J. Boley, B. W. Settlemyer and D. Katramatos, Regression-based analytics for response dynamics of SDN solutions and components, Workshop on Emerging Trends in Softwarized Networks (ETSN 2018), co-located with Netsoft2018, 2018.
4. Q. Liu, N. S. V. Rao, On concavity and utilization analytics of wide-area network transport protocols, IEEE International Conference on High Performance Computing and Communications, (HPCC), June 28-30, 2018.
5. Q. Liu, N. S. V. Rao, S. Sen, B. W. Settlemyer, H. B. Chen, J. Boley, R. Kettimuthu, and D. Katramatos, Virtual environment for testing software-defined networking solutions for scientic workows,Workshop on AI-Science -Autonomous Infrastructure for Science, in conjunction with HPDC, 2018.
6. 12. Z. liu, R. Kettimuthu, I. Foster, N. S. V. Rao, Cross-geography scientic data transfer trends and user behavior patterns, 27th ACM International Symposium on High Performance Parallel and Distributed Computing (HPDC), 2018.
7. N. S. V. Rao, C. Y. T. Ma, F. He, On defense strategies for recursive system of systems using aggregated correlations, International Conference on Information Fusion, 2018.

References

1. N. S. V. Rao, N. Imam, J. Haley, S. Oral, Wide-Area lustre le system using LNet routers, 12th Annual IEEE International Systems Conference (SYSCON2018), 2018.
2. N. S. V. Rao, C. Y. T. Ma, F. He, Defense strategies for multi-site cloud computing server infrastructures, 19th International Conference on Distributed Computing and Networking, (ICDCN 2018), 2018.
3. S. Sen, N. S. V. Rao, Q. Liu, N. Imam, I. Foster, R. Kettimuthu, Experiments and analyses of data transfers over wide-area dedicated connections, First International Workshop on Workow Science (WOWS), 2017.
4. N. S. V. Rao, Q. Liu, S. Sen, J. Hanley, I. Foster, R. Kettimuthu, C. Q. Wu, D. Yun, D. Towsley, G. Vardoyan, Experiments and analyses of data transfers over wide-area dedicated connections, The 26th International Conference on Computer Communications and Networks (ICCCN 2017), 2017.
5. N. S. V. Rao, Q. Liu, S. Sen, D. Towsley, G. Vardoyan, R. Kettimuthu, I. Foster, TCP throughput proles using measurements over dedicated connections, 26th ACM International Symposium on High Performance Parallel and Distributed Computing (HPDC), 2017.
6. N. S. V. Rao, N. Imam, C. Y. T. Ma, K. Hausken, F. He, J. Zhuang, On defense strategies forsystem of systems using aggregated correlations, 11th Annual IEEE International Systems Conference(SYSCON2017), 2017.
7. N. S. V. Rao, Q. Liu, S. Sen, G. Hinkel, N. Imam, I. Foster, R. Kettimuthu, B. Settlemyer, C. Q. Wu, D. Yun, Experimental analysis of le transfer rates over wide-area dedicated connections, 18th IEEE International Conference on High Performance Computing and Communications (HPCC), 2016.