Measuring and Characterizing the Performance of Multi-tier Cloud Applications

Ashiwan Sivakumar

Mohammad Hajjat, Shankaranarayanan P N, Sanjay Rao

LANMAN 2015

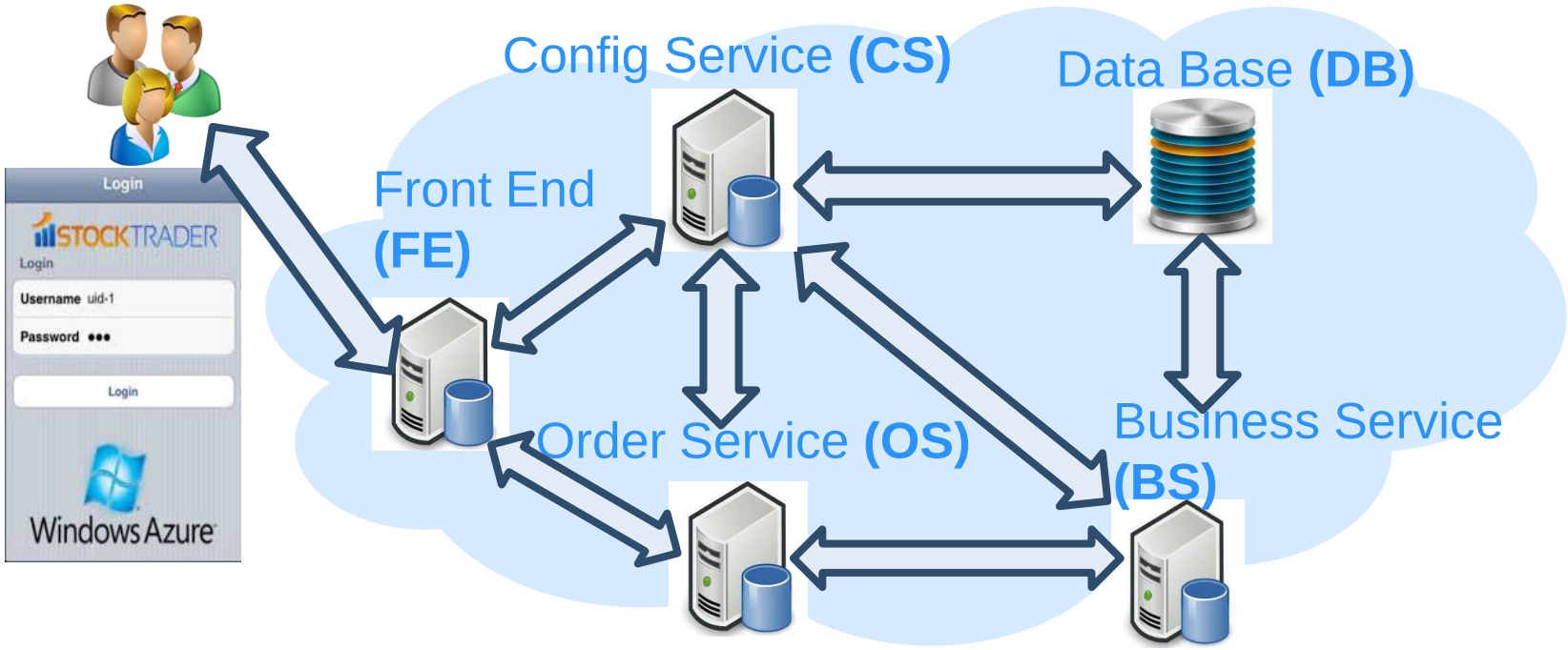
PURDUE UNIVERSITY

Motivation for our study

Interactive multi-tier applications are complex

- Multiple components with complex interactions
- Geo-distributed for high availability and low latency

E.g. Stocktrader – Components



Motivation for our study

Interactive multi-tier applications are complex

- Multiple components with complex interactions
- Geo-distributed for high availability and low latency

Require stringent SLA guarantees

- Amazon: Every 100ms costs 1% in sales
- Google: 0.5 sec delay increase → traffic and revenue drop by 20%

Cloud performance fluctuations

- **Can SLAs be met in the cloud?**

Studies characterizing performance

Existing studies :

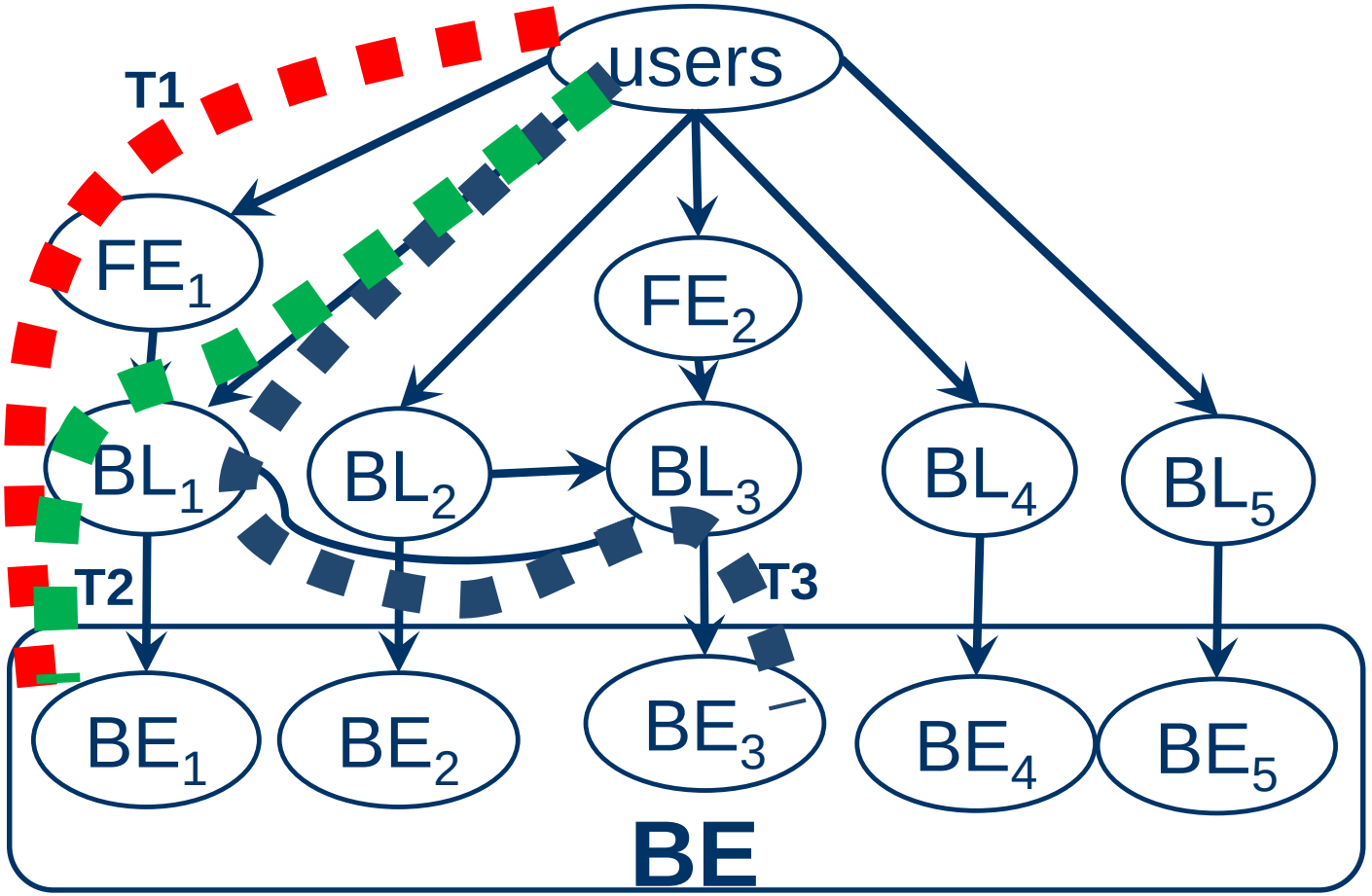
- Measure individual cloud services (E.g. EC2, Blob)
- Other classes of applications (E.g. High performance computing applications)

Our focus :

- Fine-grained per application component measurements of multi-tier apps
- Characterize performance issues experienced in the cloud

Complex transactions in geo-distributed multi-tier apps

E.g. Thumbnail, Stocktrader, ERP

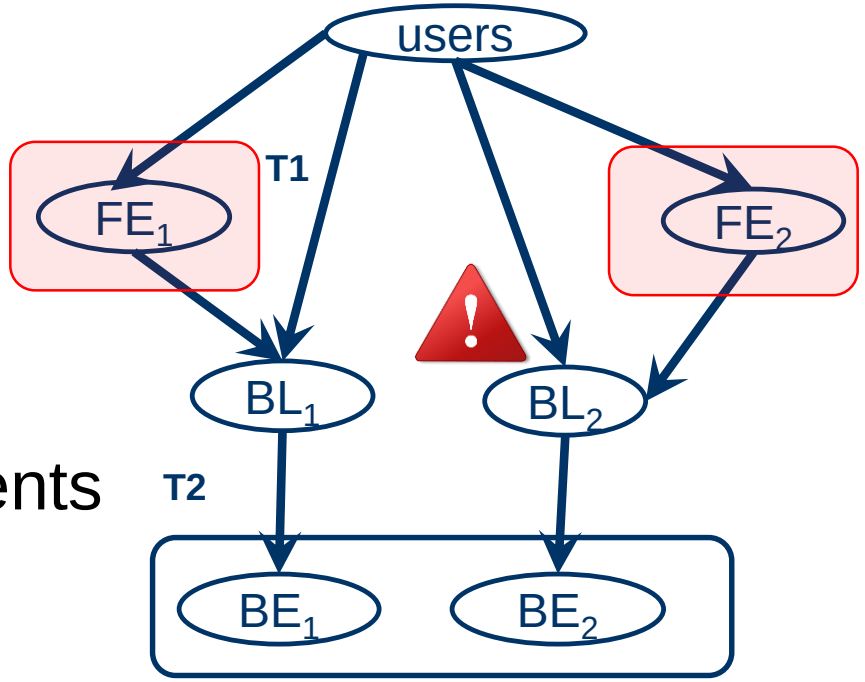


Our Contributions

- Characterization of perf.

in a geo-distributed setting

- Per-component measurements



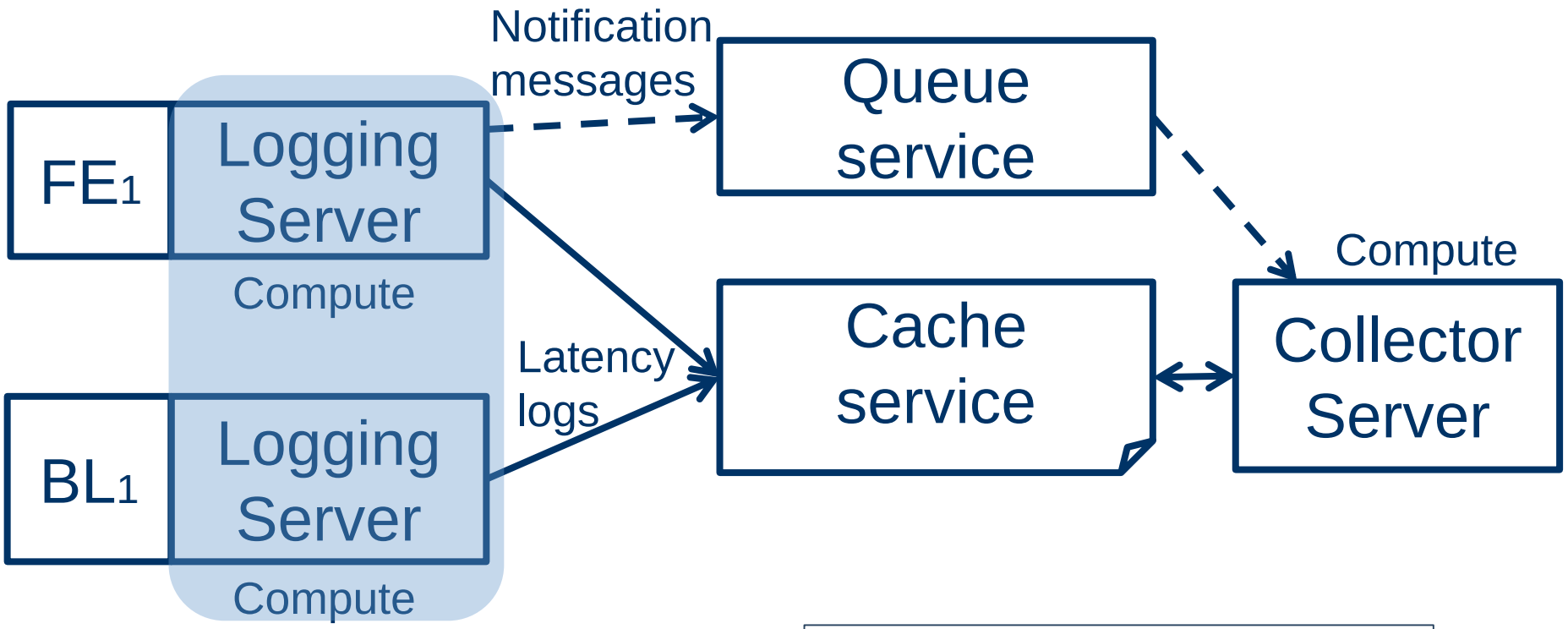
Key findings :

- **Replicas of a component are uncorrelated** across **DCs**
- Attributed to a **few app components** **at any time**
- Performance issues are **short-lived; 90% < 4 mins**
- **Choosing the best replica combination across DCs** gives higher latency reduction

Outline

- Monitoring framework & Evaluation setup
- Characterization of poor performance
- Exploiting geo-distribution
- Conclusions

Monitoring framework

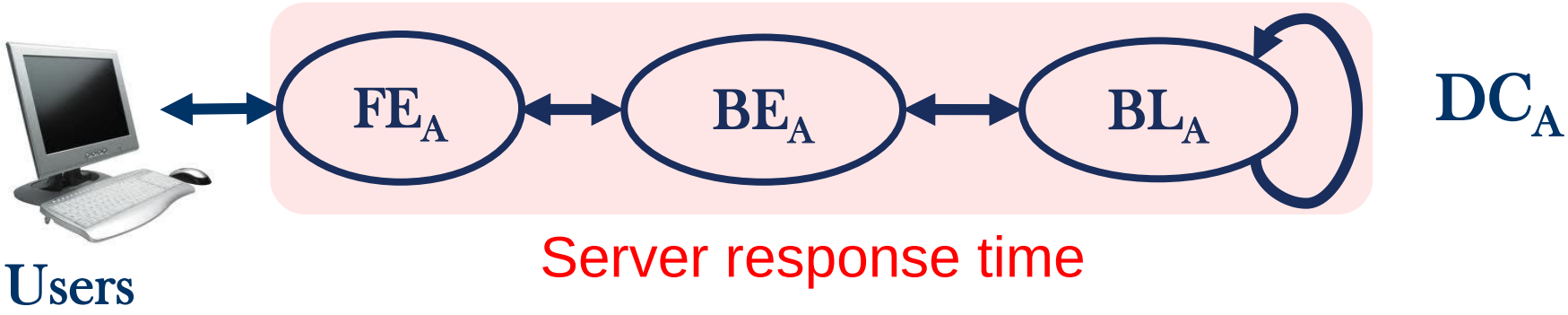


X-Trace + aspects (AOP) for portability

Per-component measurement tools – “not general” (Dapper) and “not portable” (X-Trace)

Evaluation setup

- **Two cloud platforms** - Microsoft Azure, Amazon AWS
- **Four Applications**
  - Data-intensive : Thumbnail
  - Delay-sensitive : Stocktrader, Daytrader
  - Social : Twissandra
- **Real benchmark** workload (E.g. DaCapo)
- Metric – **server response time** (no internet delay)



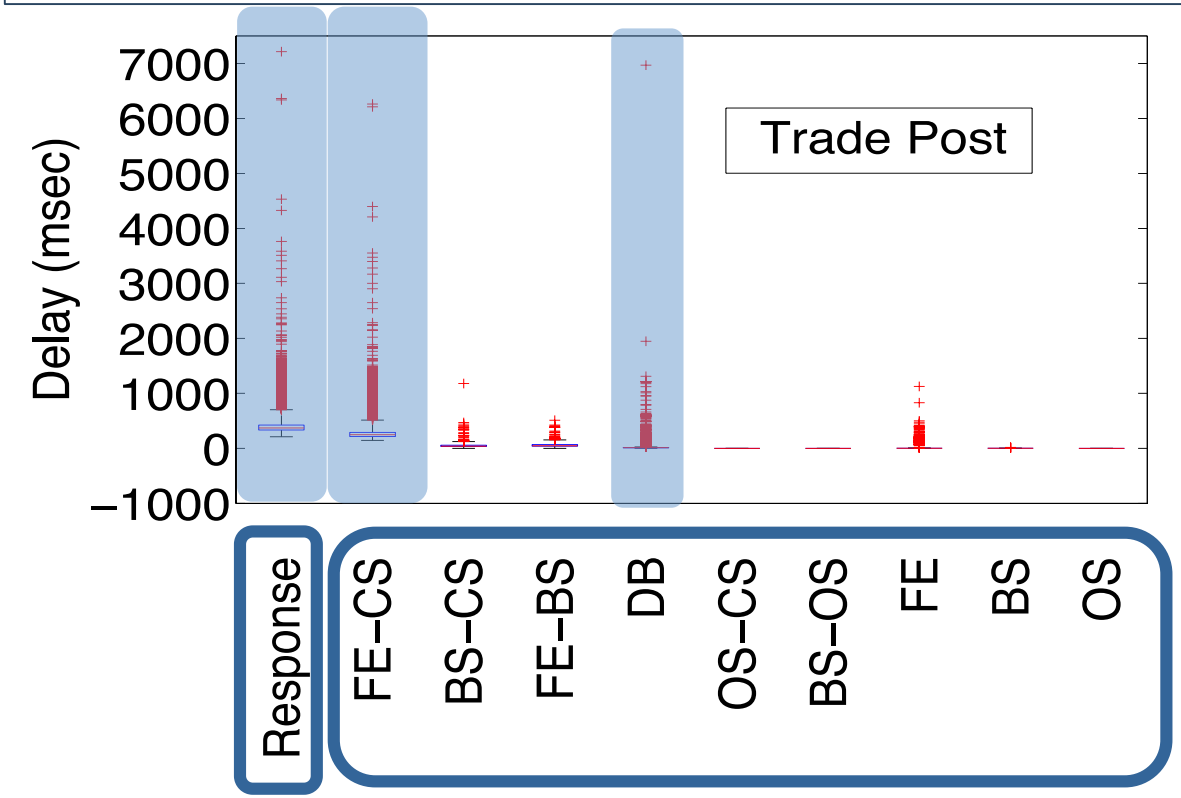
Outline

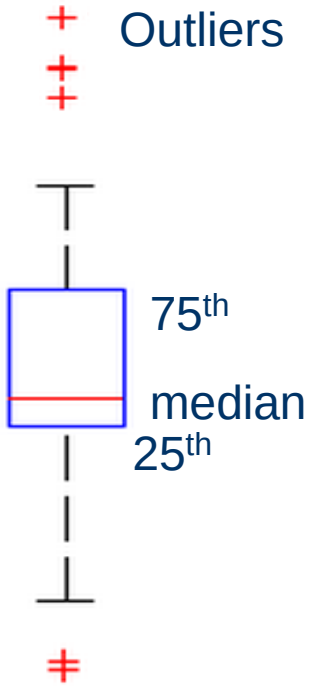
- Monitoring framework & Evaluation setup
- Characterization of poor performance
- Exploiting geo-distribution
- Conclusions

Dissecting performance into constituent components

Long tail and Variation in all transactions

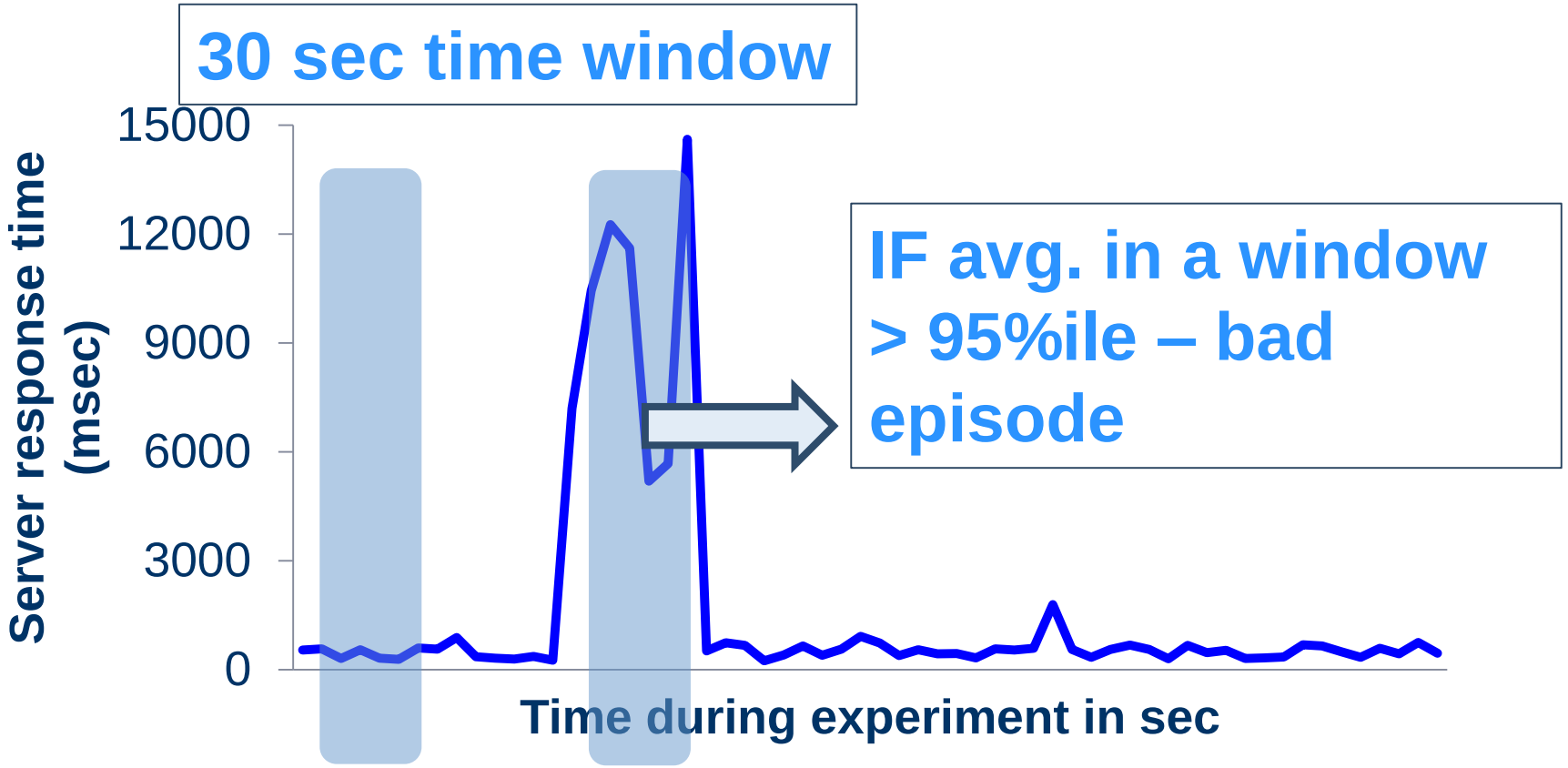
E.g. Simple login - 99.9%ile/median is 28



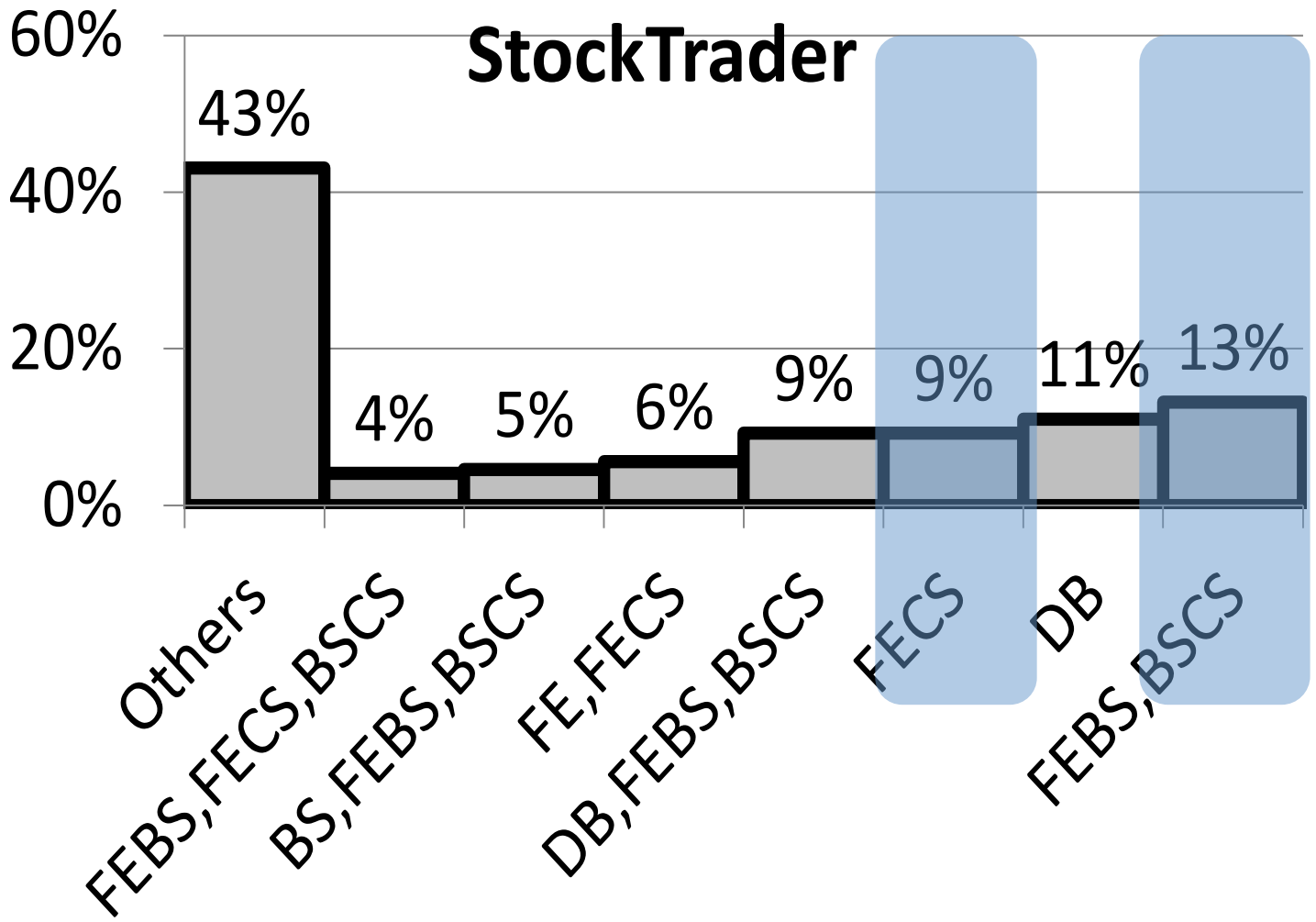


Few components show more variation

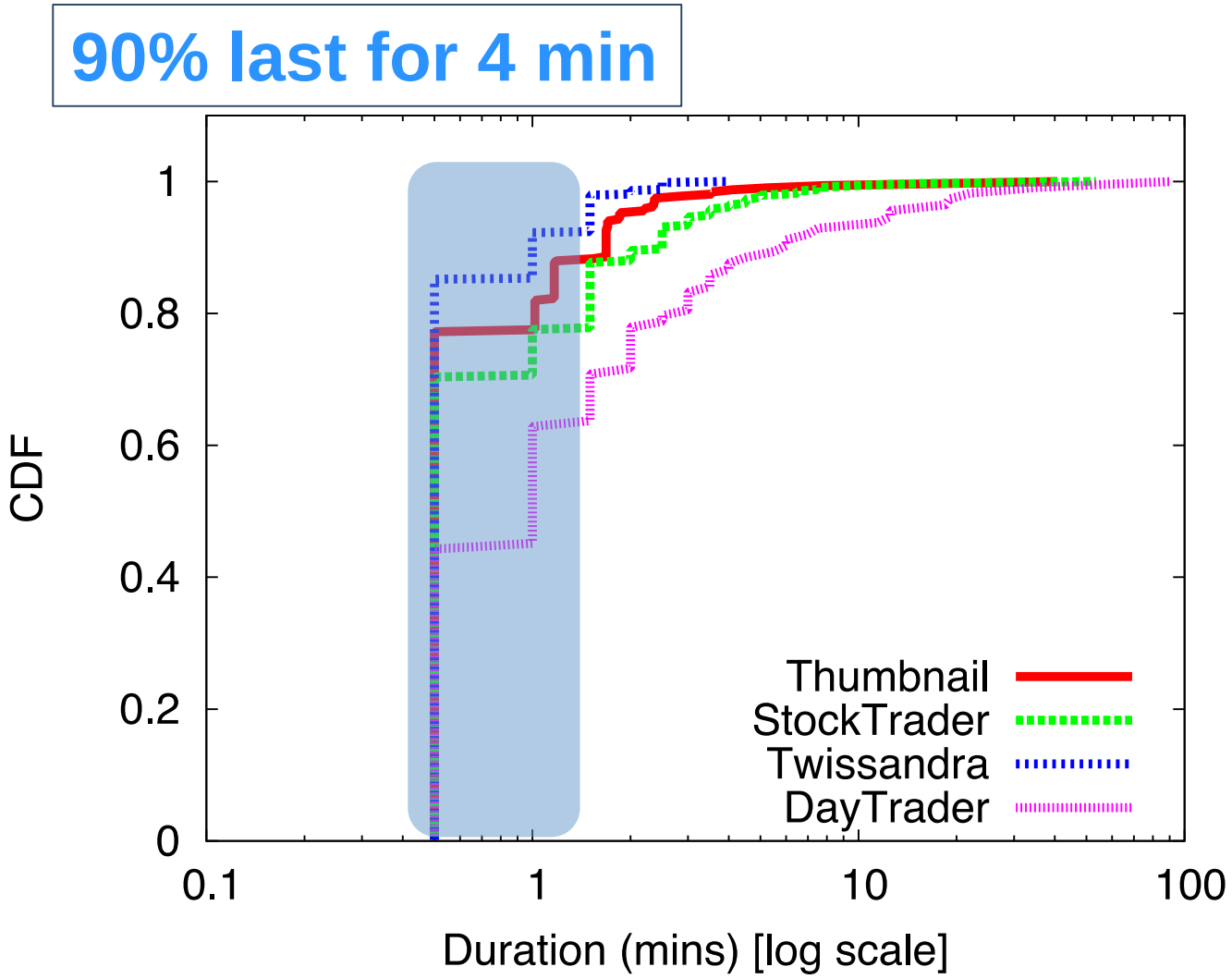
Analyzing bad performance episodes



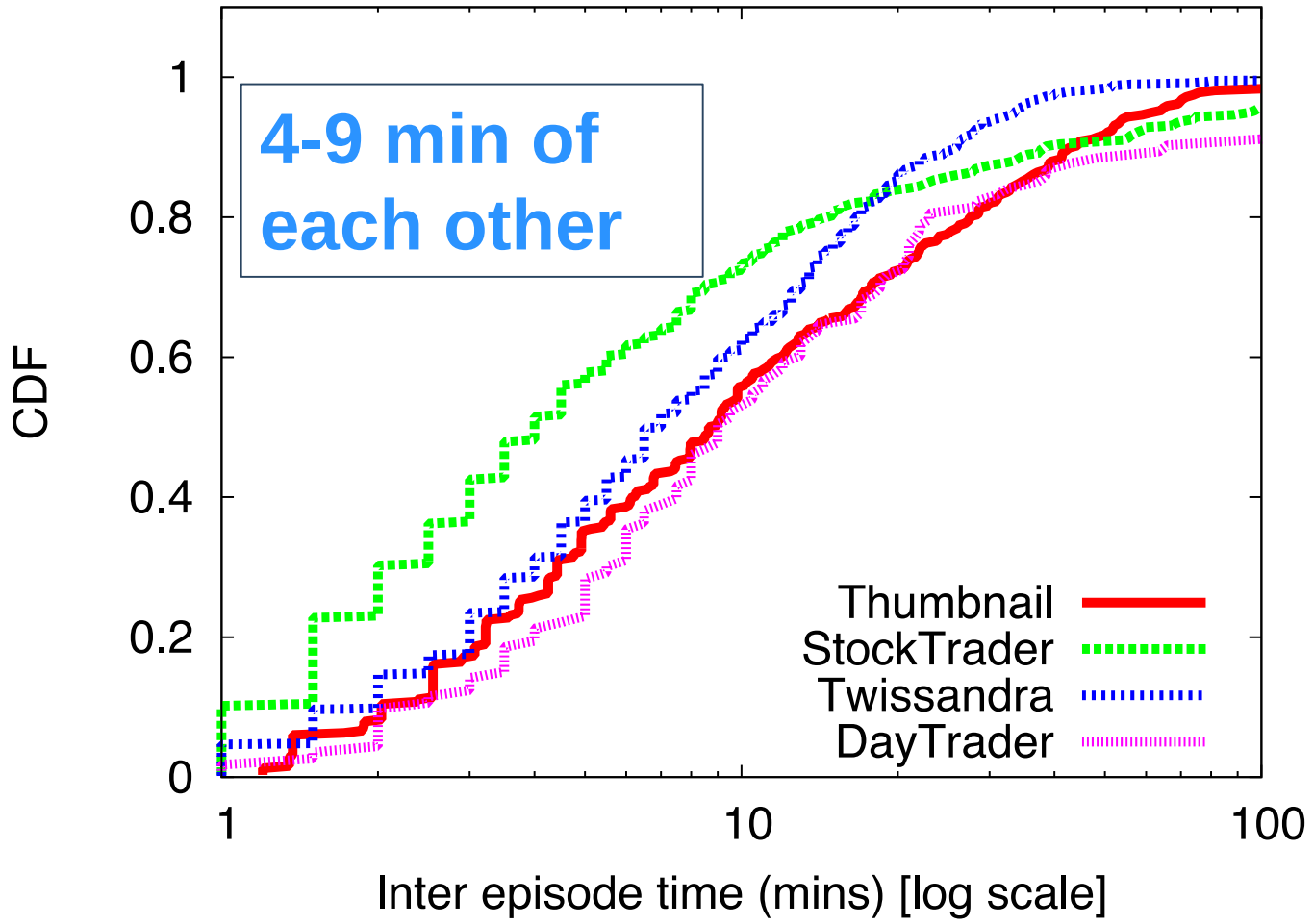
Different subset responsible for bad performance at different times



Bad performance episodes are short-lived

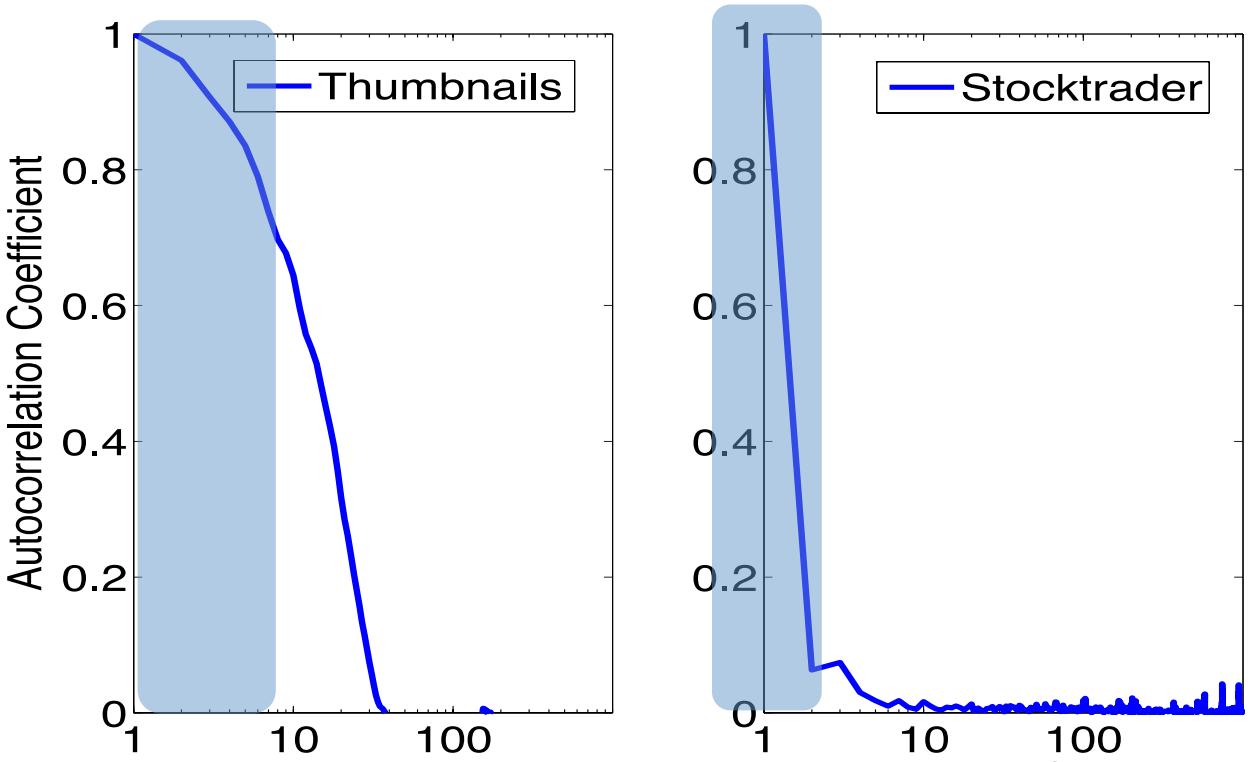


Bad performance episodes occur frequently



Persistence of performance

- Auto-correlation function measure
  - Tendency for “server response time” to remain in the same state over time



Predictable over short time (<10 min)

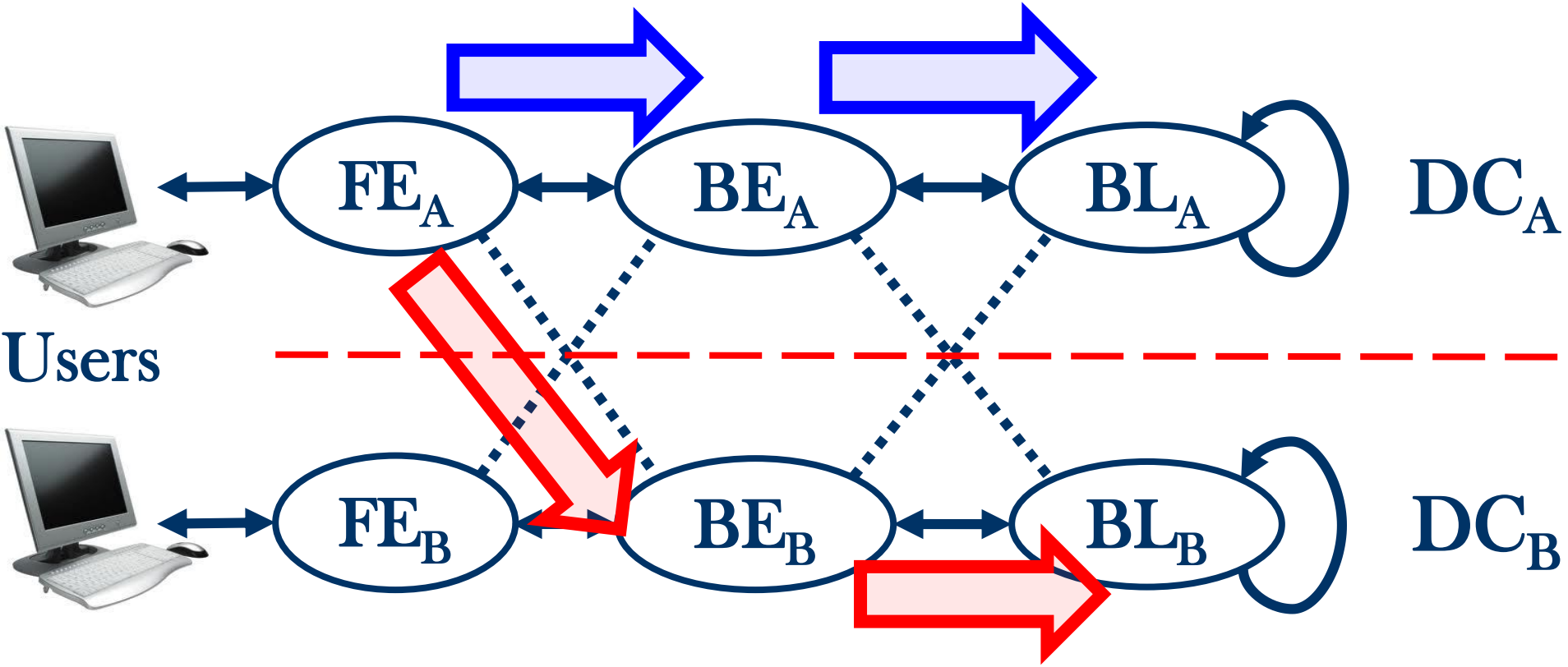
Other result

- Performance of component replicas across DCs
  - Uncorrelated

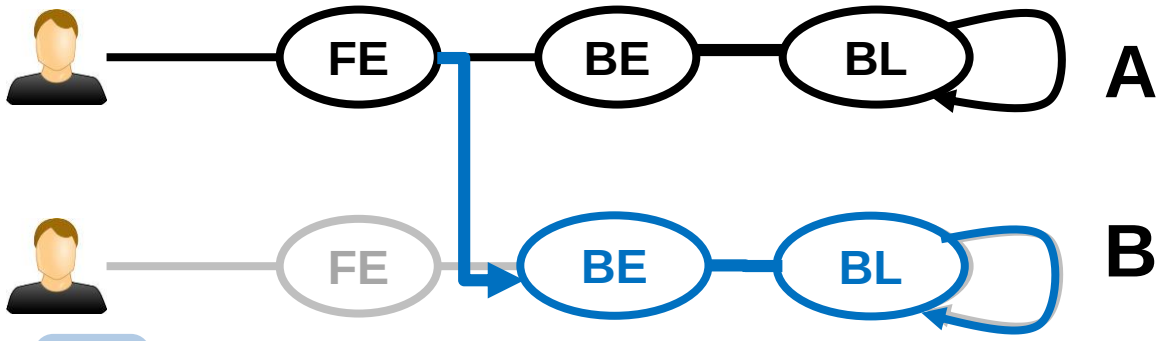
Outline

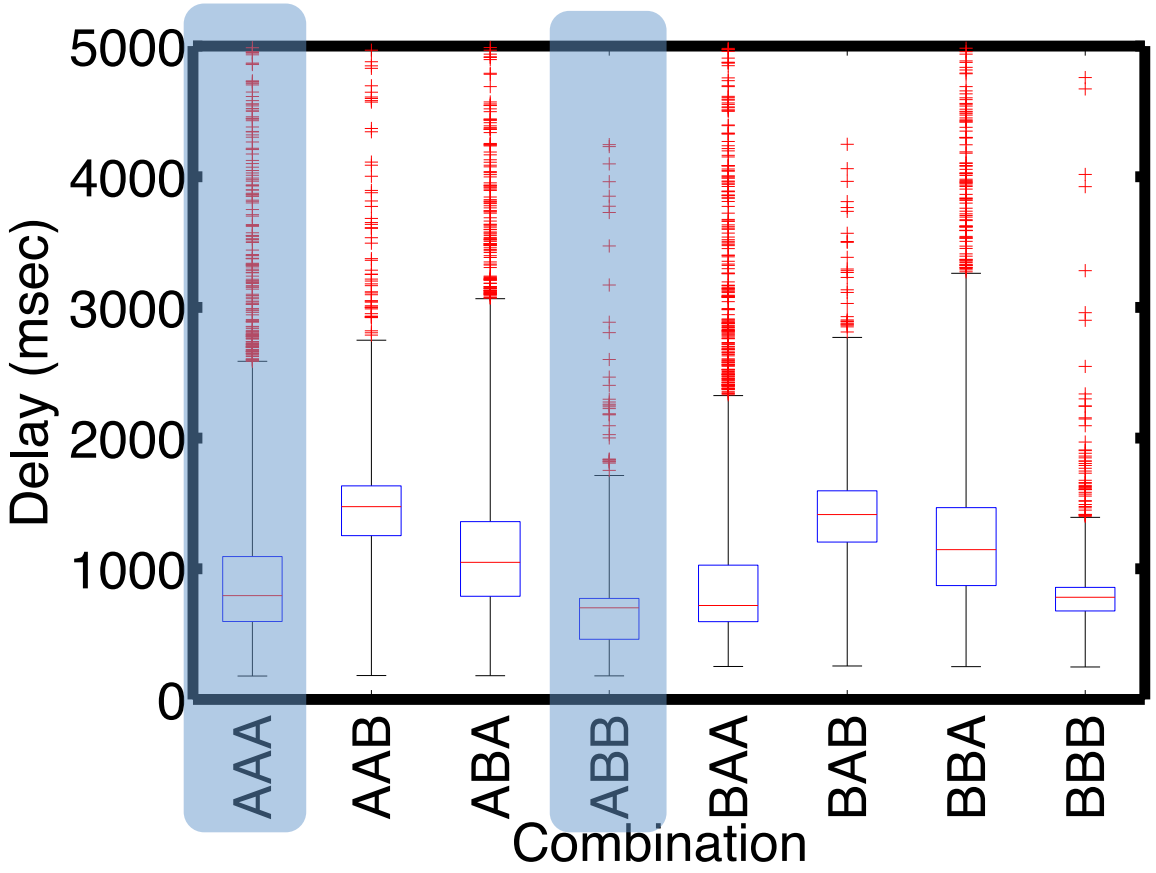
- Monitoring framework & Evaluation setup
- Characterization of poor performance
- Exploiting geo-distribution
- Conclusions

Exploiting geo-distribution



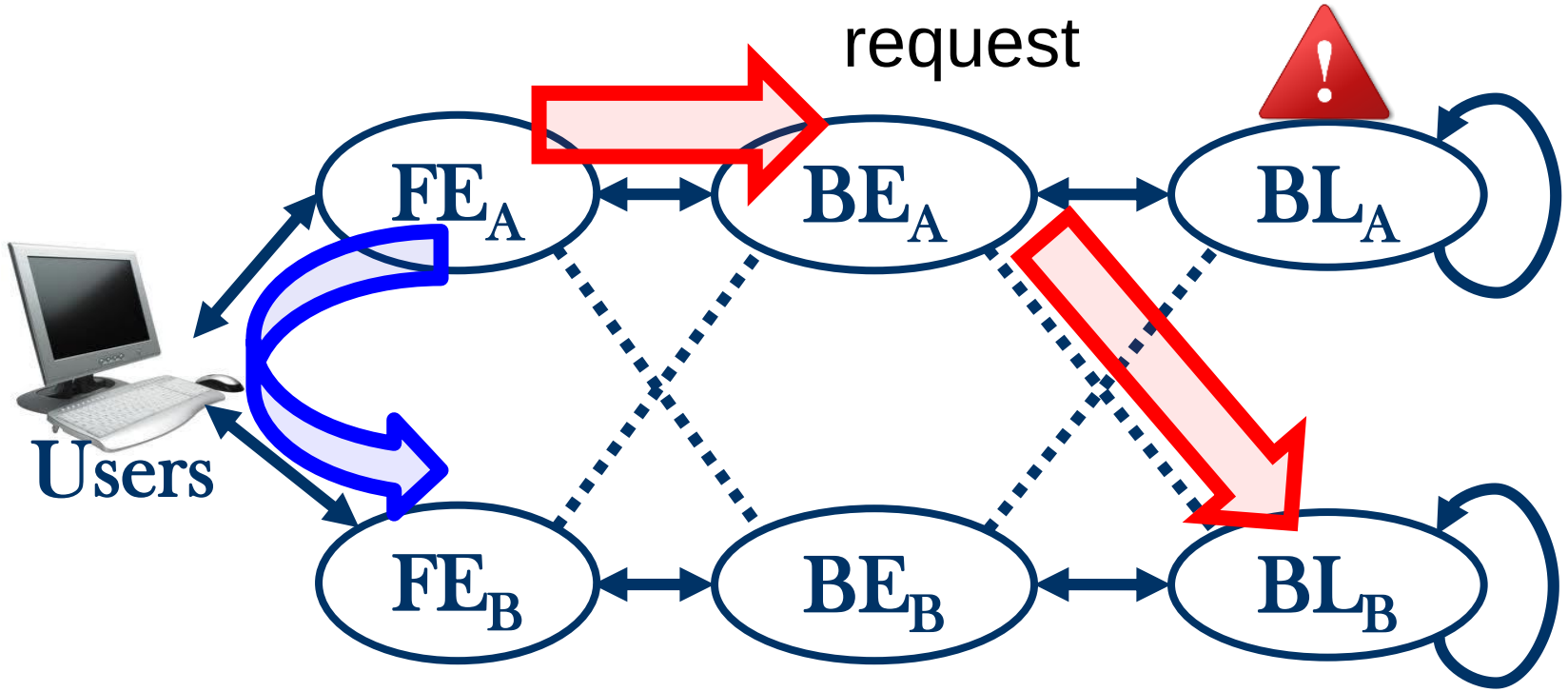
Cross DC path performs better sometimes



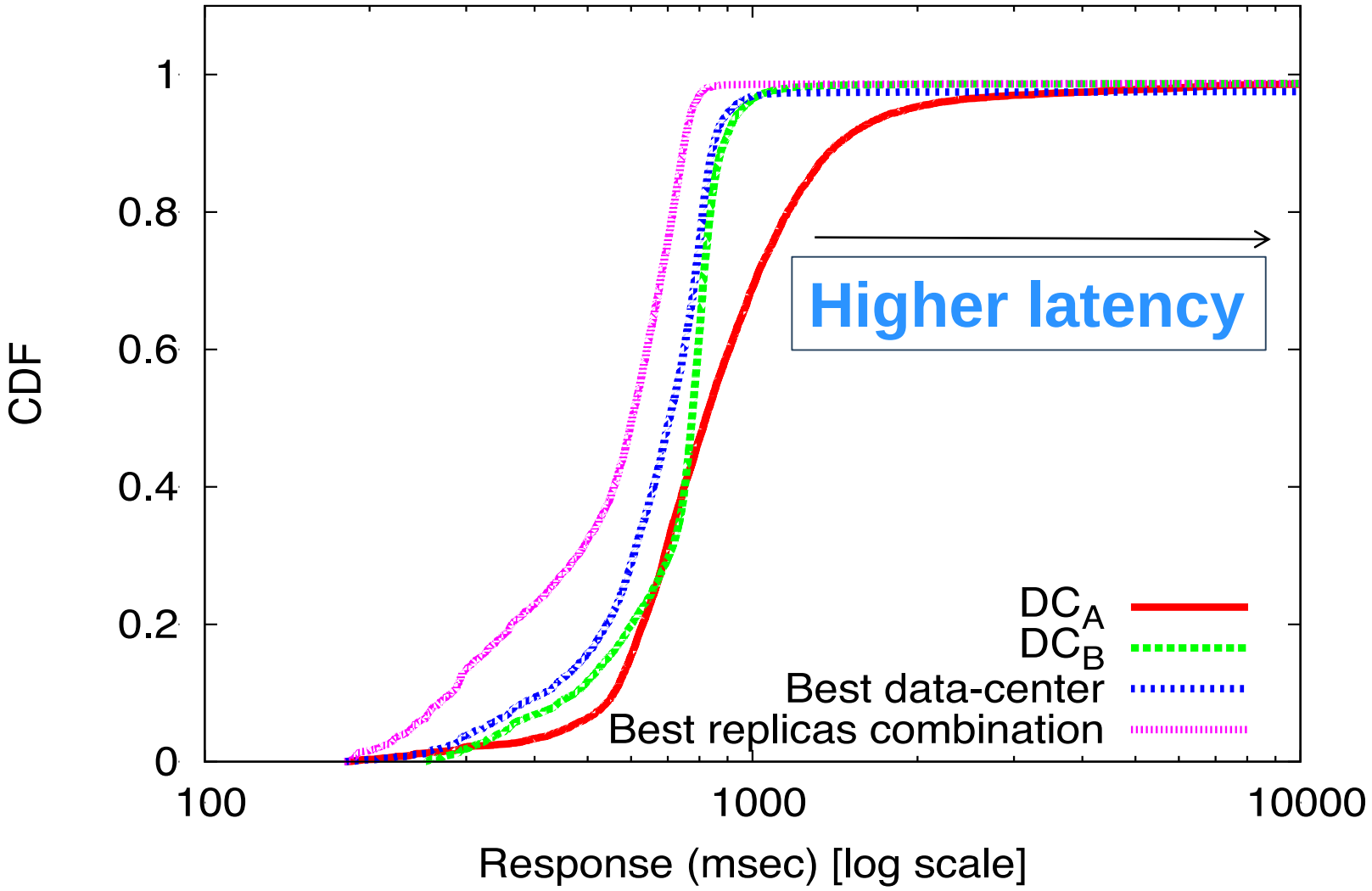


Best DC Vs. Best replica redirection strategies

- **Best DC :** Re-route entire request at the granularity of DCs
- **Best replica :** Select the best replica combination for each request



Best replica combination gives best results



Dealer: per-component request splitting

- **Dealer:** handle cloud variability in multi-tier interactive apps [CoNEXT 2012, JSAC 2013]
- Per-component re-routing: dynamically split user req's across replicas in multiple DC's at component granularity
- Transient cloud variability: performance problems in cloud services, workload spikes, failures, etc.
- Performance tail improvement:
  - Natural cloud dynamics > 6x

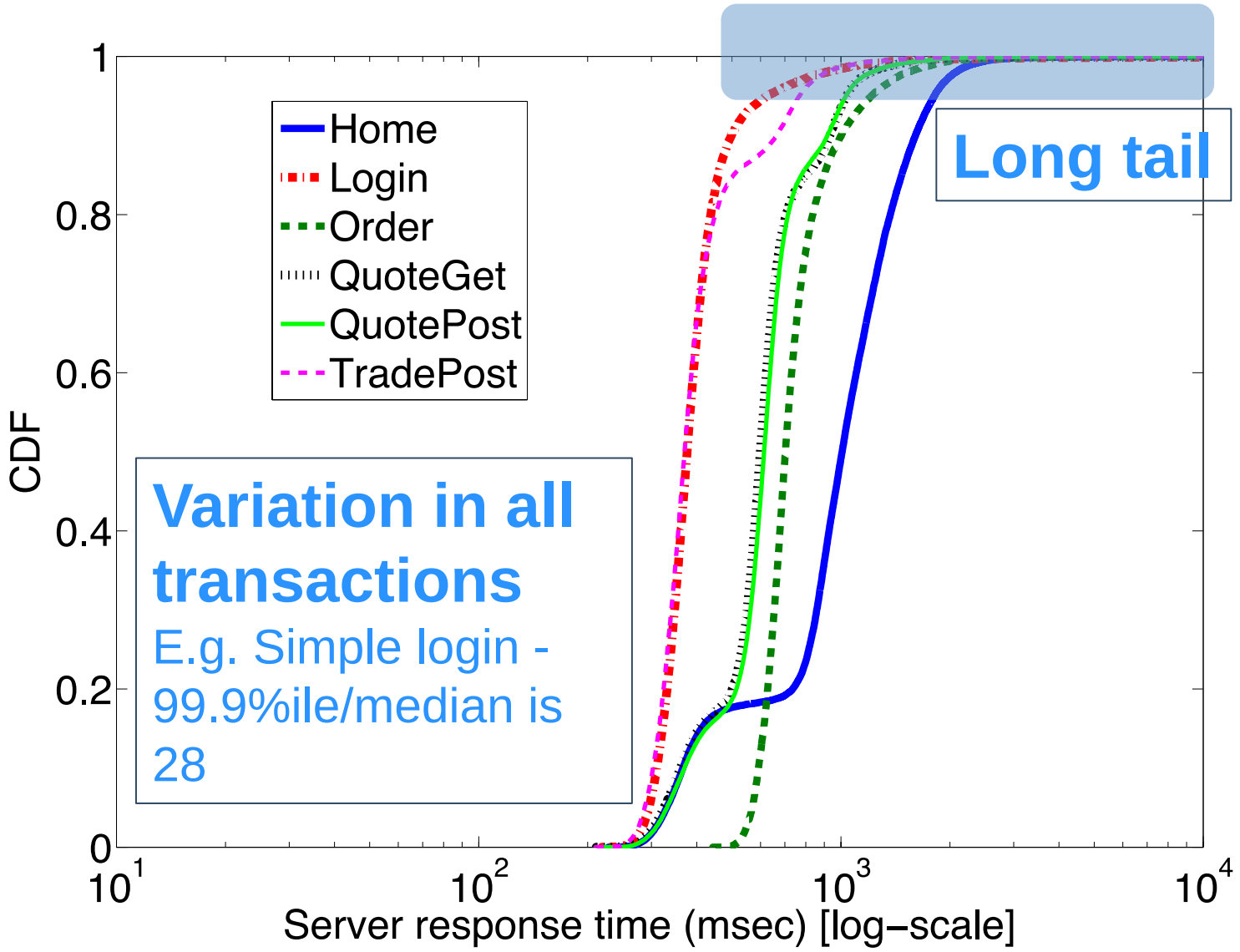
Conclusions

- Presented a performance characterization of multi-tier apps in the cloud
  - To answer the question – Can SLA guarantees be met in the cloud?
- Applications experience short-term performance fluctuations frequently attributed to a few app components in a subset of DCs
- Choosing the best replica combination across DCs gives higher latency reduction than coarse-grained strategies

Q&A

Backup

Performance by transaction type - Stocktrader



Correlation coefficients

	FE	DB	BS	OS	FE-BS	FE-CS	BS-CS	BS-OS	OS-CS
FE	1	-0.08	-0.11	-0.04	-0.31	0.03	-0.32	-0.07	-0.04
DB	1	0.50	0.03	-0.01	-0.01	0.04	0.05	0.02	
BS	1	0.14	0.08	-0.02	0.09	0.14	0.14		
OS	1	-0.37	-0.03	-0.40	0.66	0.74			
FE-BS	1	0.01	0.87	-0.31	-0.37				
FE-CS	1	-0.01	-0.02	-0.03					
BS-CS	1	-0.34	-0.41						
BS-OS	1	0.71							
OS-CS	1								

(a) StockTrader