Workload-driven Analysis of File Systems in Shared Multi-Tier Data-Centers over InfiniBand

K. Vaidyanathan P. Balaji H. -W. Jin D.K. Panda

Network-Based Computing Laboratory

Department of Computer Science and Engineering

The Ohio State University

Presentation Outline

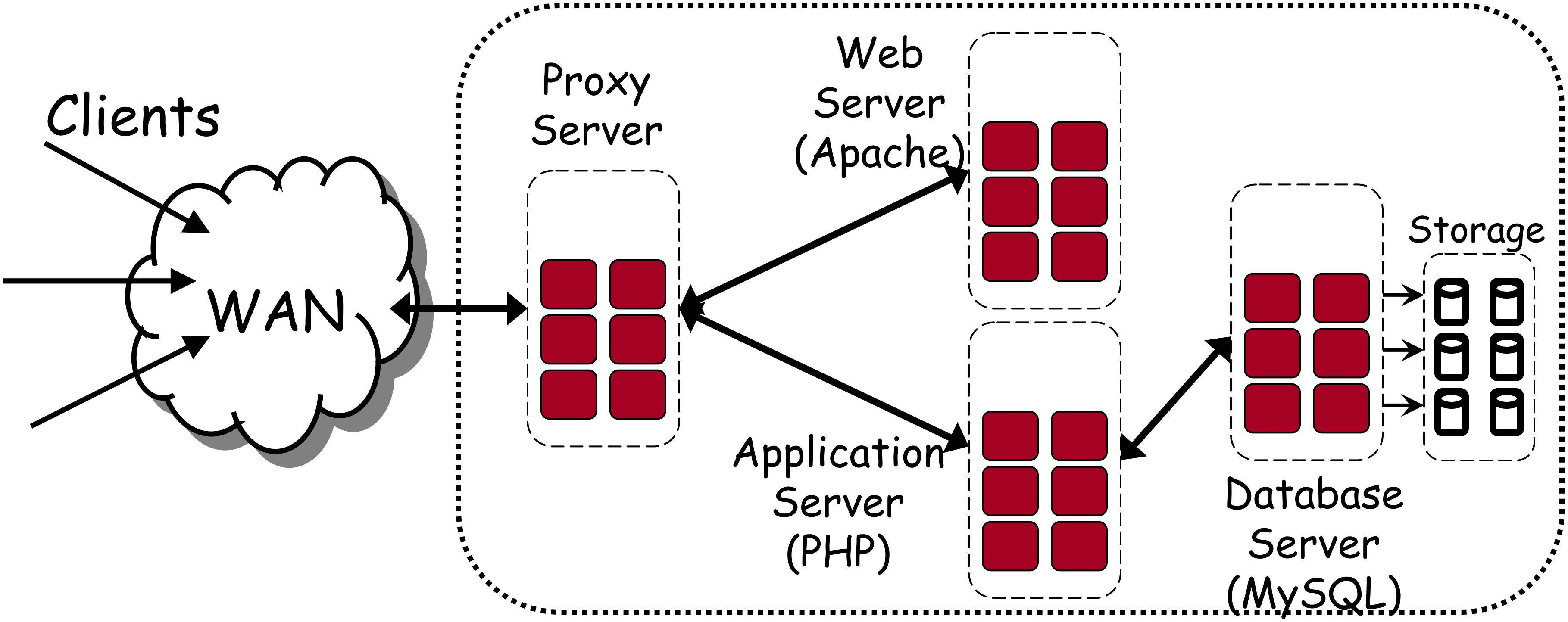
- **Introduction and Background**
- Characterization of local and network-based file systems
- Multi File System for Data-Centers
- Experimental Results
- Conclusions

Introduction

- Exponential growth of Internet
  - Primary means of electronic interaction
  - Online book-stores, World-cup scores, Stock markets
  - Ex. Google, Amazon, etc
- Highly Scalable and Available Web-Services
- **Performance is critical for such Services**
- Utilizing Clusters for Web-Services? [shah01]
  - High Performance-to-cost ratio
  - Has been proposed by Industry and Research Environments

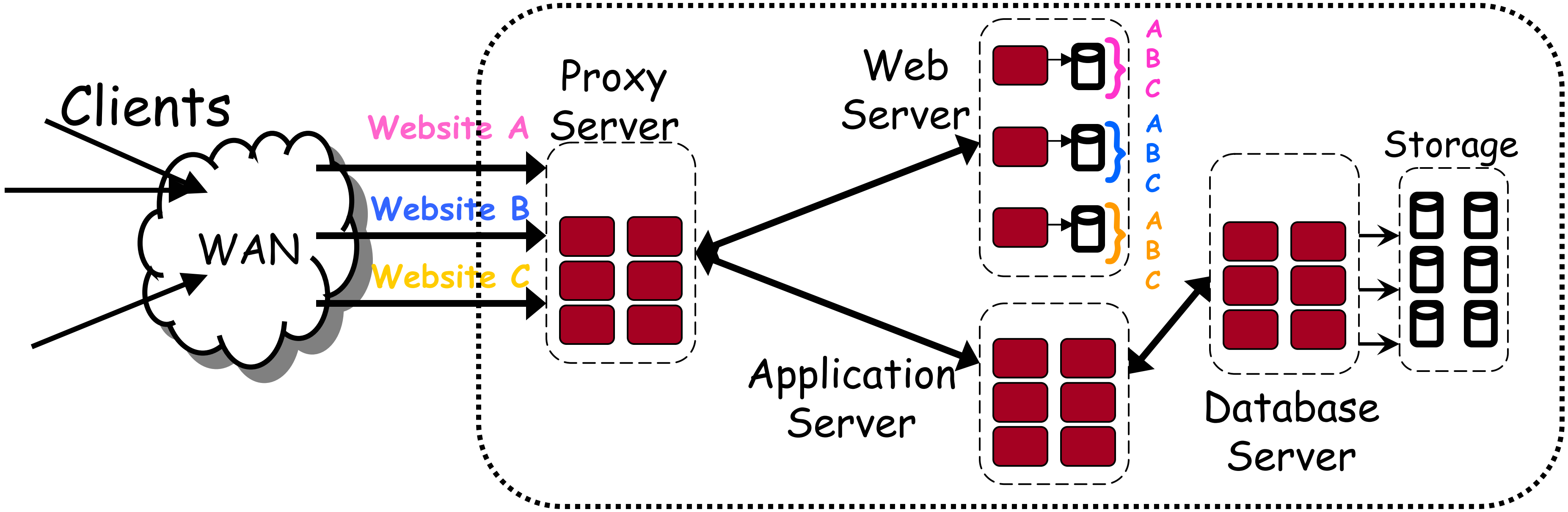
[shah01]: CSP: A Novel System Architecture for Scalable Internet and Communication Services. H. V. Shah, D. B. Minturn, A. Foong, G. L. McAlpine, R. S. Madukkarumukumana and G. J. Regnier In USITS 2001

Cluster-Based Data-Centers



- Nodes are logically partitioned
  - provides specific services (serving static and dynamic content)
  - Use high speed interconnects like InfiniBand, Myrinet, etc.
- Requests get forwarded through multiple tiers
- Replication of content on all nodes

Shared Cluster-Based Data-Centers

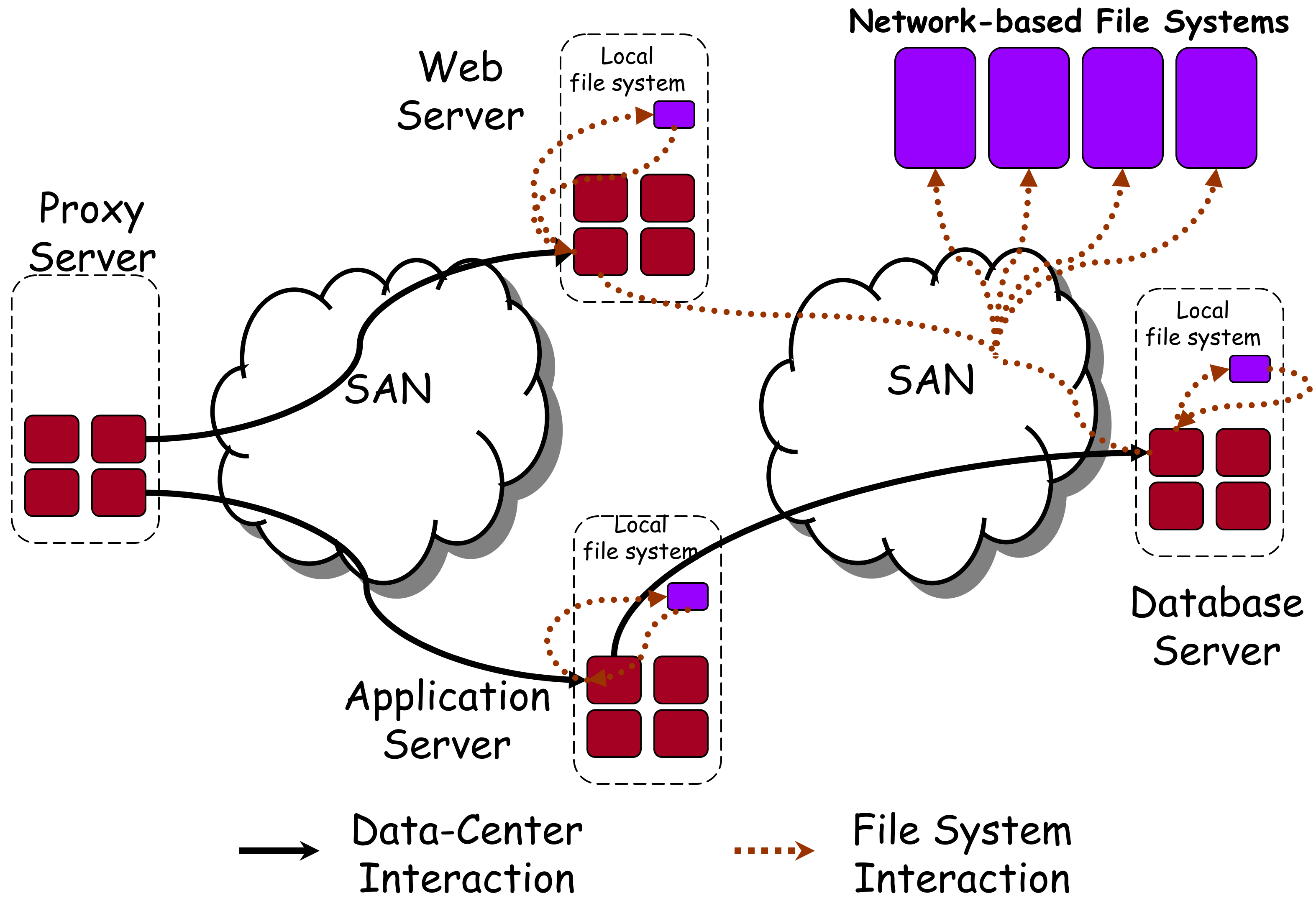


- Hosting several unrelated services on a single data-center
  - Currently used by several ISPs and Web Service Providers (IBM, HP)
- Replication of content
  - *Amount of data replicated increases linearly with the number of web-sites hosted*

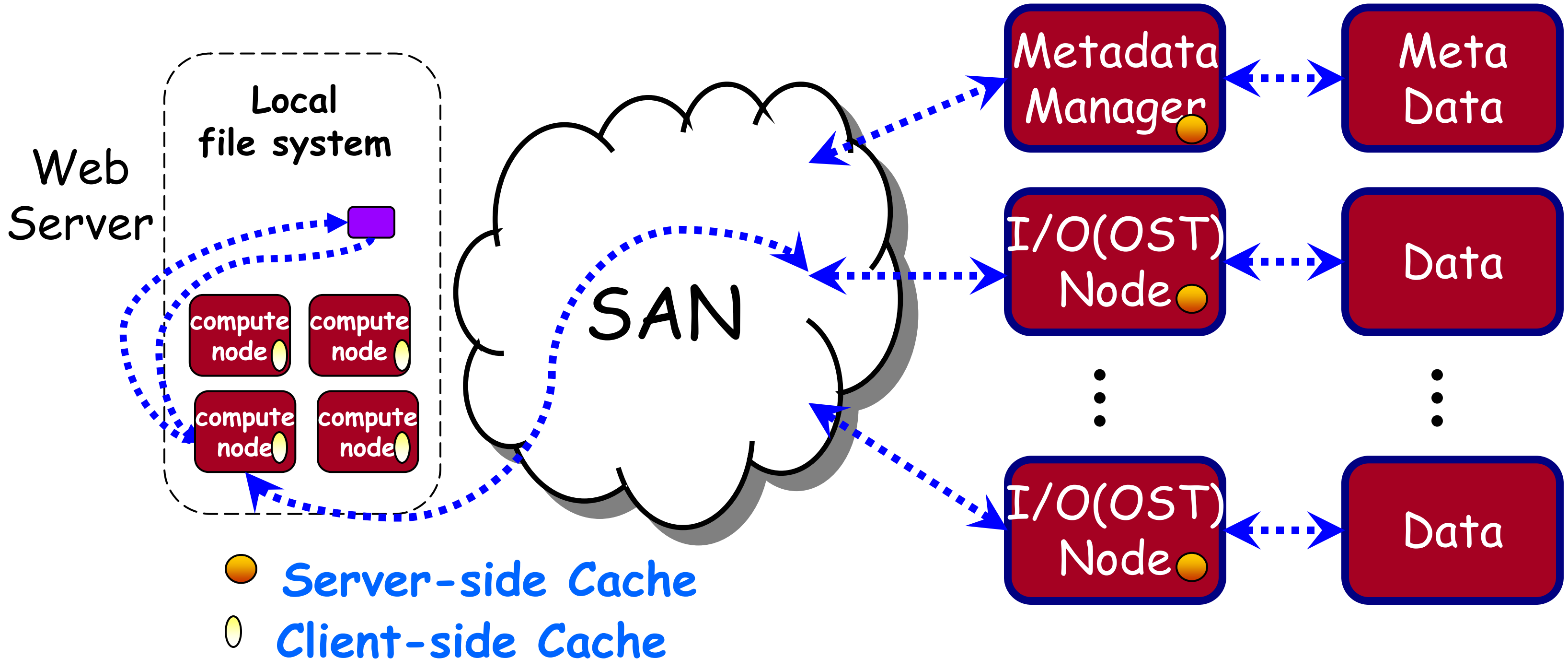
Issues in Shared Cluster-Based Data-Centers

- File System Caches being shared across multiple web-sites
- Under-utilization of aggregate cache of all nodes
- Web-site Content
  - Replication of content on all nodes if we use local file system
  - Need to fetch the document via network if we use network file system, however no replication required
- **Can we adapt the file system to avoid these?**

File System Interactions



Existing File Systems



- Network-based File System: Parallel Virtual File System (PVFS) and Lustre (supports client-side caching)
- Local File System: ext3fs and memory file system (ramfs)

Presentation Outline

- Introduction and Background
- Characterization of local and network-based file systems
- Multi File System for Data-Centers
- Experimental Analysis
- Conclusions

Characterization of local and network-based File Systems

- Network Traffic Requirements
- Aggregate Cache
- Cache Pollution Effects

Network Traffic Requirements

- Absolute Network Traffic generated
  - Static Content
  - Dynamic Content
- Network Utilization
  - Large/Small burst (static or dynamic content)
- Overhead of Metadata Operations

Aggregate Cache in Data-Centers

- Local File Systems use only single node's cache
  - Small files get huge benefits, if in memory. Otherwise, we pay a penalty of accessing the disk
  - Large Files may not fit in memory and also have high penalties in accessing the disk
- Network File Systems use aggregate cache from all nodes
  - Large Files, if striped, can reside in file system cache on multiple nodes
  - Small files also get benefits due to aggregate cache

Cache Pollution Effects

- Working set – frequently accessed documents; usually fits in memory
- Shared Data-Centers
  - Multiple web-sites share the file system cache; each website has lesser amount of file system cache to utilize
  - Bursts of requests/accesses to one web-site may result in cache pollution
  - May result in drastic drop in the number of cache hits

Presentation Outline

- Introduction and Background
- Characterization of local and network-based file systems
- **Multi File System for Data-Centers**
- Experimental Results
- Conclusions

Multi File System for Data-Centers

Characterization	ext3fs	ramfs	pvfs	lustre
Network Traffic generated	Min	Min	More traffic	Min
Use of Aggregate Cache	No	No	Yes	Yes
Cache pollution effects	Yes	No	Yes	Yes
Metadata overhead	No	No	Yes	Yes

Multi File System for Data-Centers

- A combination of file systems for different environments
- Memory file system and local file system (ext3fs) for workloads with high temporal locality
- Memory file system and network file system (pvfs/lustre) for workloads with low temporal locality

Presentation Outline

- Introduction and Background
- Characterization of local and network-based file systems with data-centers
- Multi File System for Data-Centers
- **Experimental Results**
- Conclusions

Experimental Test-bed

- Cluster 1 with:
  - 8 SuperMicro SUPER X5DL8-GG nodes; Dual Intel Xeon 3.0 GHz processors
  - 512 KB L2 Cache, 2 GB memory; PCI-X 64 bit 133 MHz
- Cluster 2 with:
  - 8 SuperMicro SUPER P4DL6 nodes; Dual Intel Xeon 2.4 GHz processors
  - 512 KB L2 Cache, 512 MB memory; PCI-X 64 bit 133 MHz
- Mellanox MT23108 Dual Port 4x HCAs; MT43132 24-port switch
- Apache 2.0.48 Web and PHP 4.3.7 Servers; MySQL 4.0.12, PVFS 1.6.2, Lustre 1.0.4

Workloads

- Zipf workloads: the relative probability of a request for the i^{th} most popular document is proportional to $1/i^{\alpha}$ with $\alpha \leq 1$
  - High Temporal locality (constant α)
  - Low Temporal locality (varying α)
- TPC-W traces according to the specifications

Class	File Sizes	Size
Class 0	1K - 250K	25 MB
Class 1	1K - 1MB	100 MB
Class 2	1K - 4MB	450 MB
Class 3	1K - 16MB	2 GB
Class 4	1K - 64MB	6 GB

Experimental Analysis (Outline)

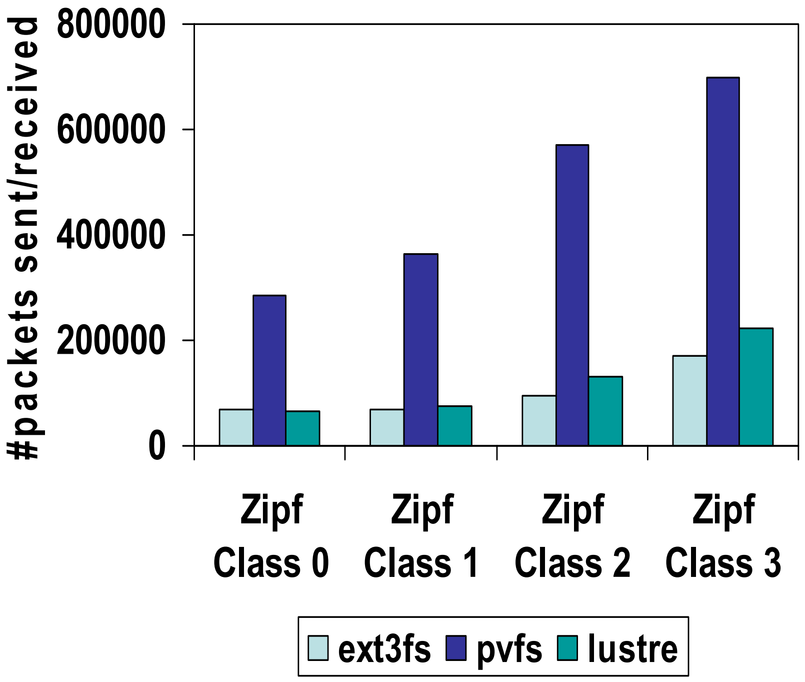
- Basic Performance of different file systems
- Network Traffic Requirements
- Impact of Aggregate Cache
- Cache Pollution Effects
- Multi File System for Data-Centers

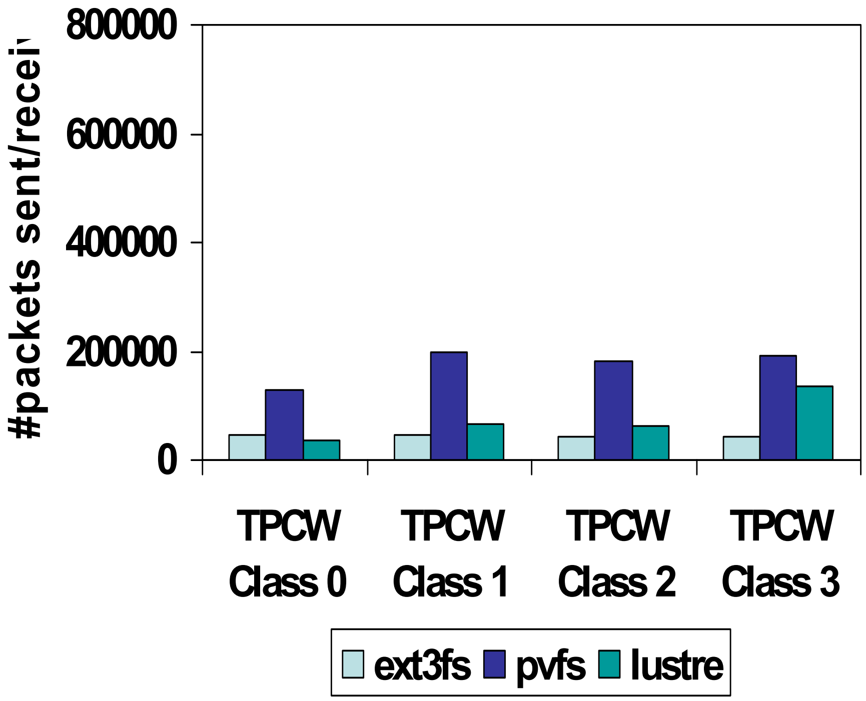
Basic Performance

Latency	ext3fs (usecs)		ramfs (usecs)		pvfs (usecs)		lustre (usecs)	
	4K	1M	4K	1M	4K	1M	4K	1M
Open & Close overhead	6	6	6	6	**1060**	**1060**	**876**	**876**
Read Latency (cache)	4	1602	4	1578	680	13825	**7.7**	1998
Read Latency (no cache)	1500	76312	1400	2379	9600	44108	3000	50713

- Network File Systems incur **high overhead for metadata operations** (open() and close())
- Lustre supports client-side cache
- For large files, network-based file system does better than local file system due to striping of the file

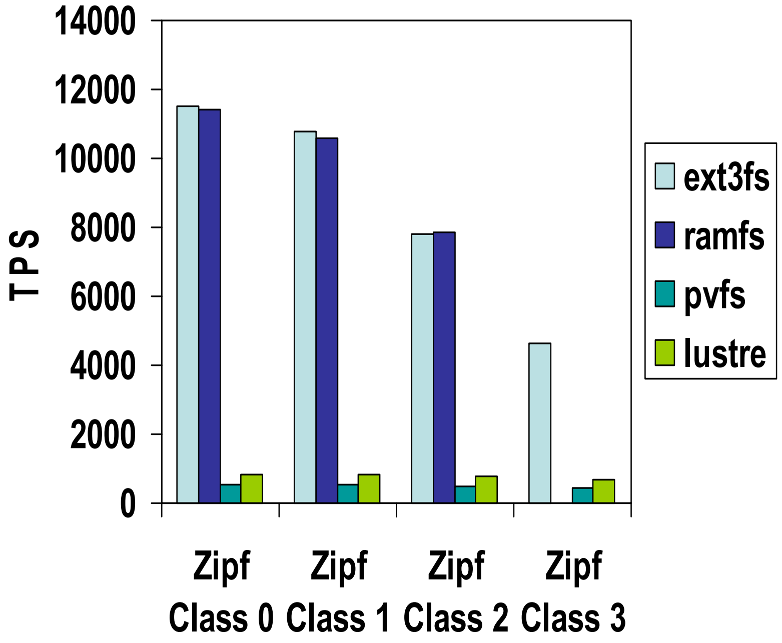
Network Traffic Requirements

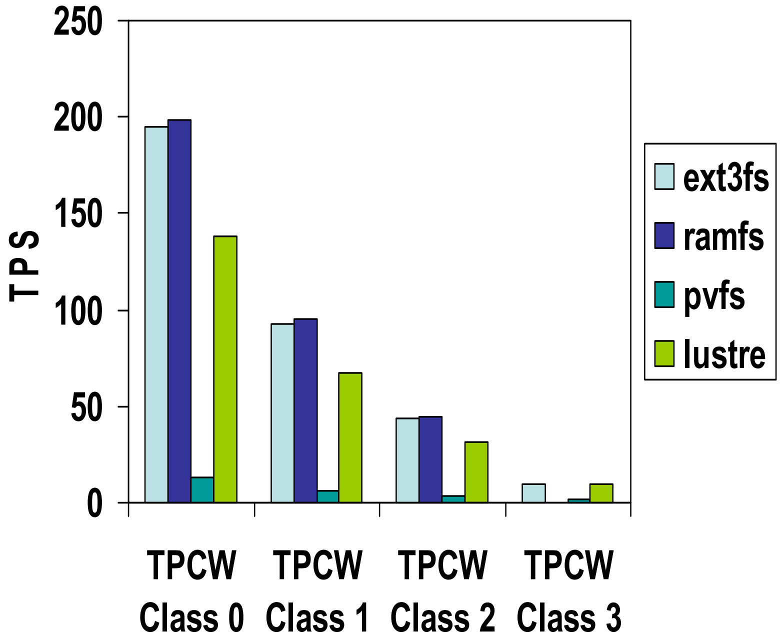




- Absolute Network Traffic Generated:
  - Increases proportionally compared to the local file system for PVFS
  - For Lustre, the traffic is close to that of the local file system
  - For dynamic content, the network traffic does not increase with increase in database size

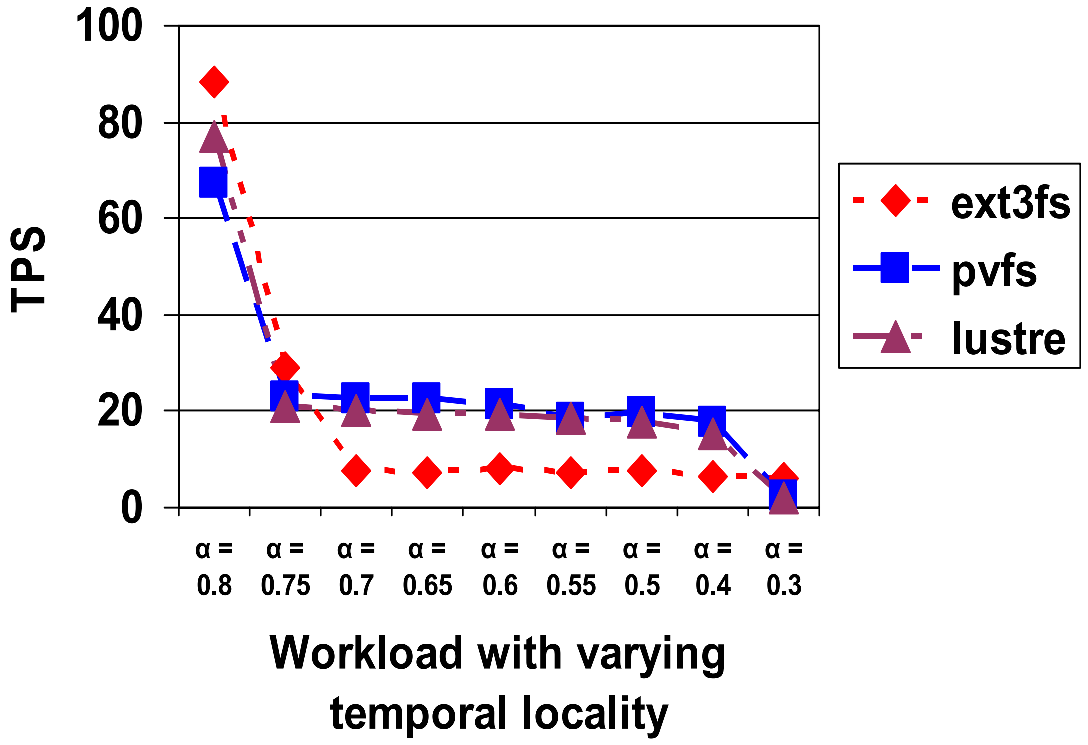
Impact of Caching and Metadata operations





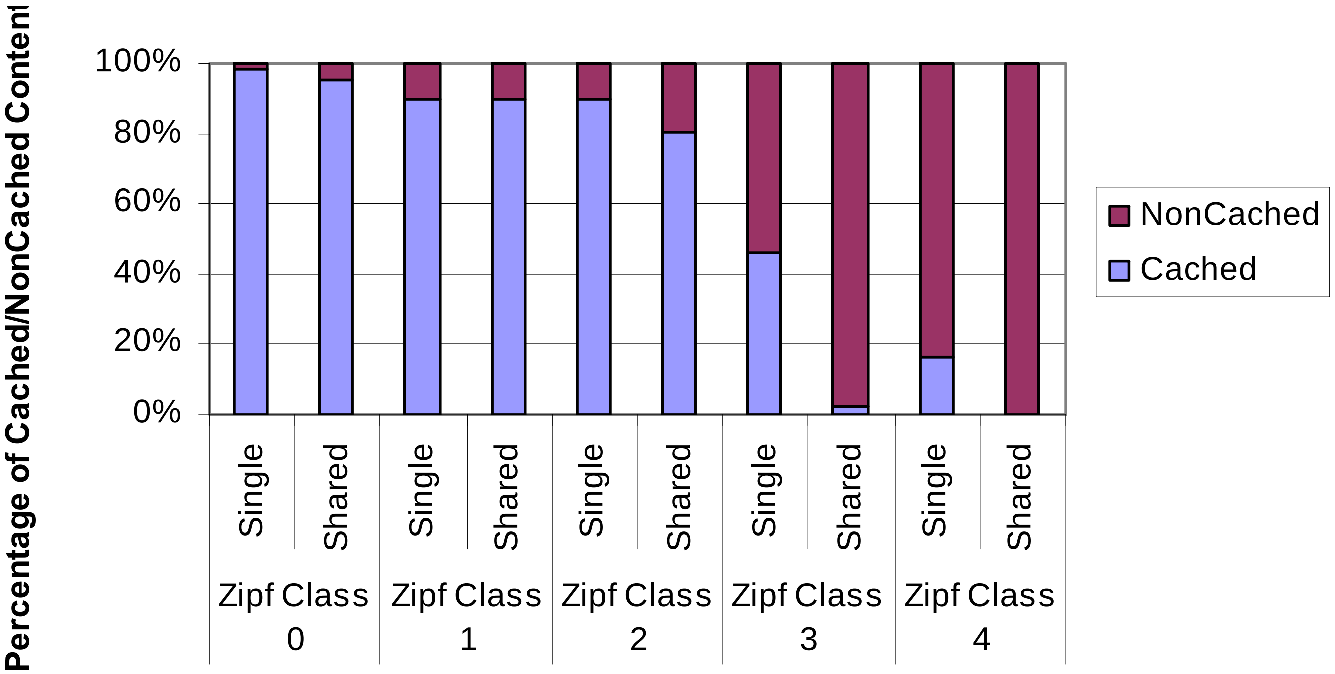
- Local File Systems are better for workloads with high temporal locality
- Surprisingly Lustre performs comparable with local file systems

Impact of Aggregate Cache



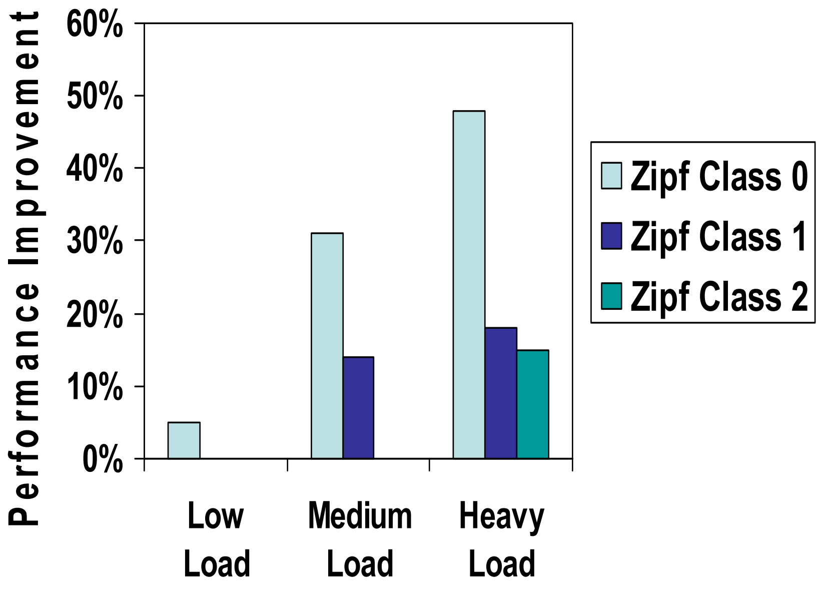
- Aggregate Cache improves data-center performance for network-based file systems

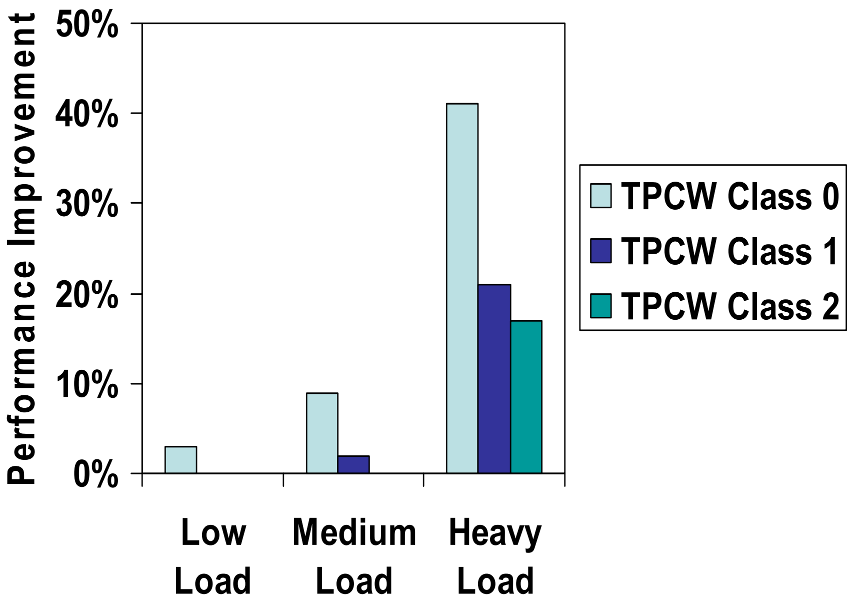
Cache Pollution Effects in Shared Data-Centers



- Small Workloads, web-sites are not affected
- Large Workloads, cache pollution affects multiple web-sites
- Placing files on memory file system might avoid the cache pollution effects

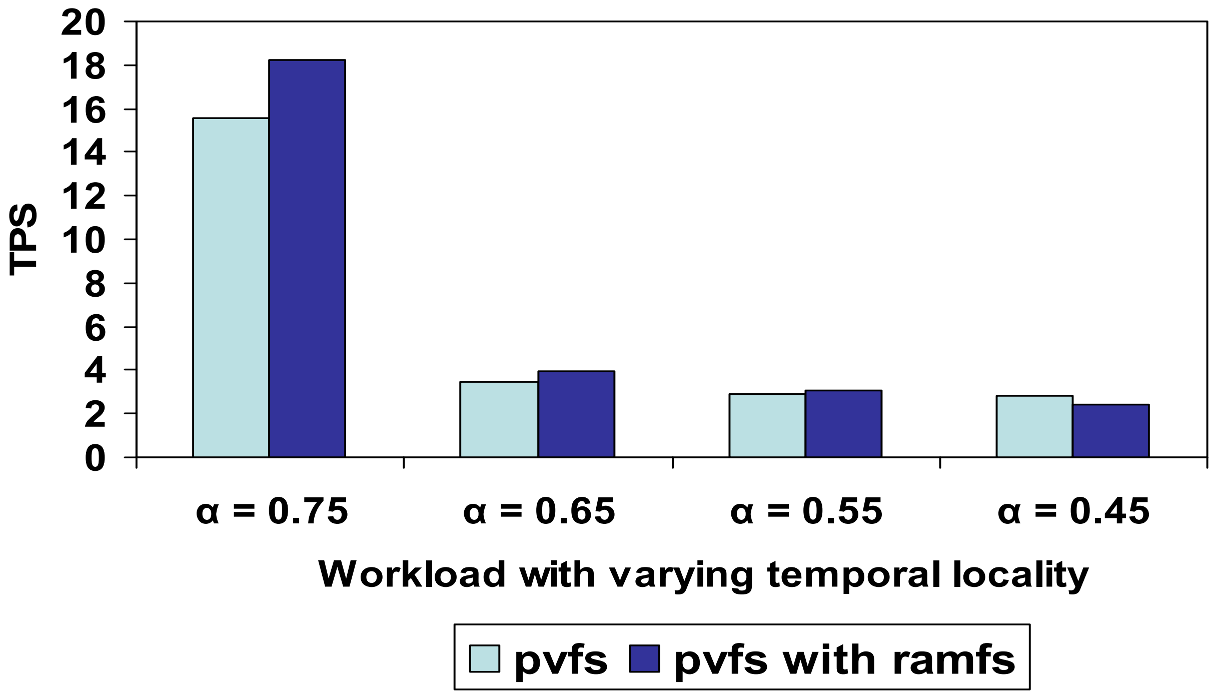
Multi File System Data-Centers





- Performance benefits for static content is close to 48%
- Performance benefits for dynamic content is close to 41%

Multi File System Data-Centers



- Benefits are two folds:
  - Avoidance of Cache Pollution
  - Reduced overhead of open() and close() operations for small files

Conclusions & Future Work

- Fragmentation of resources in shared data-Centers
  - Under-utilization of file system cache in clusters
  - Cache Pollution affects performance
- Studied the impact of file systems in terms of network traffic, aggregate cache and cache pollution effects
- Proposed a Multi File System approach to utilize the benefits from each file system
  - Combination of Network and Memory File System for static content with low temporal locality
  - Memory File System and local file system for static content with high temporal locality and dynamic content
- Propose to perform dynamic reconfiguration based on each node's memory cache and provide prioritization and QoS

Web Pointers

NOWLAB

http://www.cse.ohio-state.edu/~panda

http://nowlab.cse.ohio-state.edu

{vaidyana,balaji,jinhy,panda}@cse.ohio-state.edu