

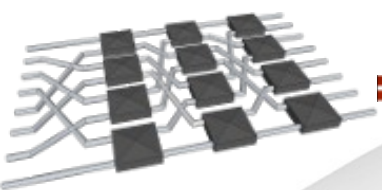
[Interaction of Access Patterns on the dNFSp File System]

Rodrigo Kassick, Francieli Zanon,
Philippe Navaux



GPPD



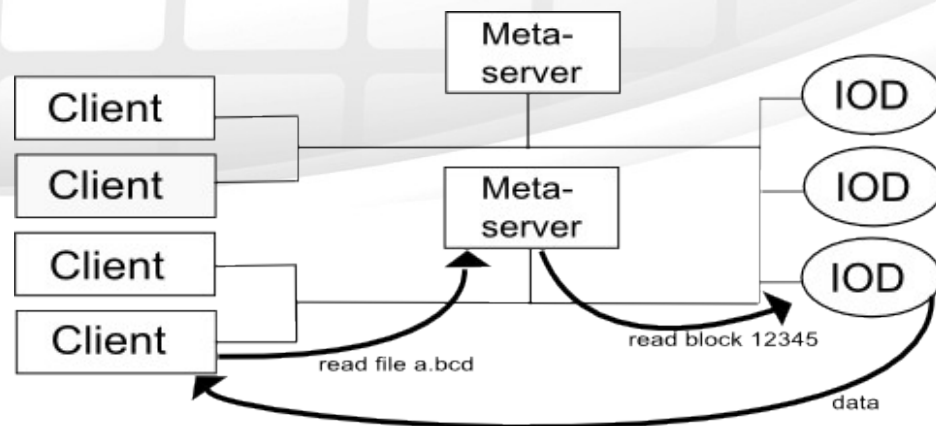


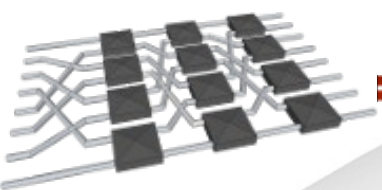
HPC Applications

- HPC: Distributed Applications running on hundreds of Processors
- Great amount of Data
 - Needs to be available as input to execution nodes
 - Data generated as the result of simulations needs to be available after the execution
- Need for a **high capacity** and **scalable** storage infra-structure

Parallel File System: dNFSp

- Distributes data over Set of Servers – IOD's
- NFS Protocol between clients and *Meta-Servers*
 - Distributed meta-data service
 - Proxies requests to IOD's

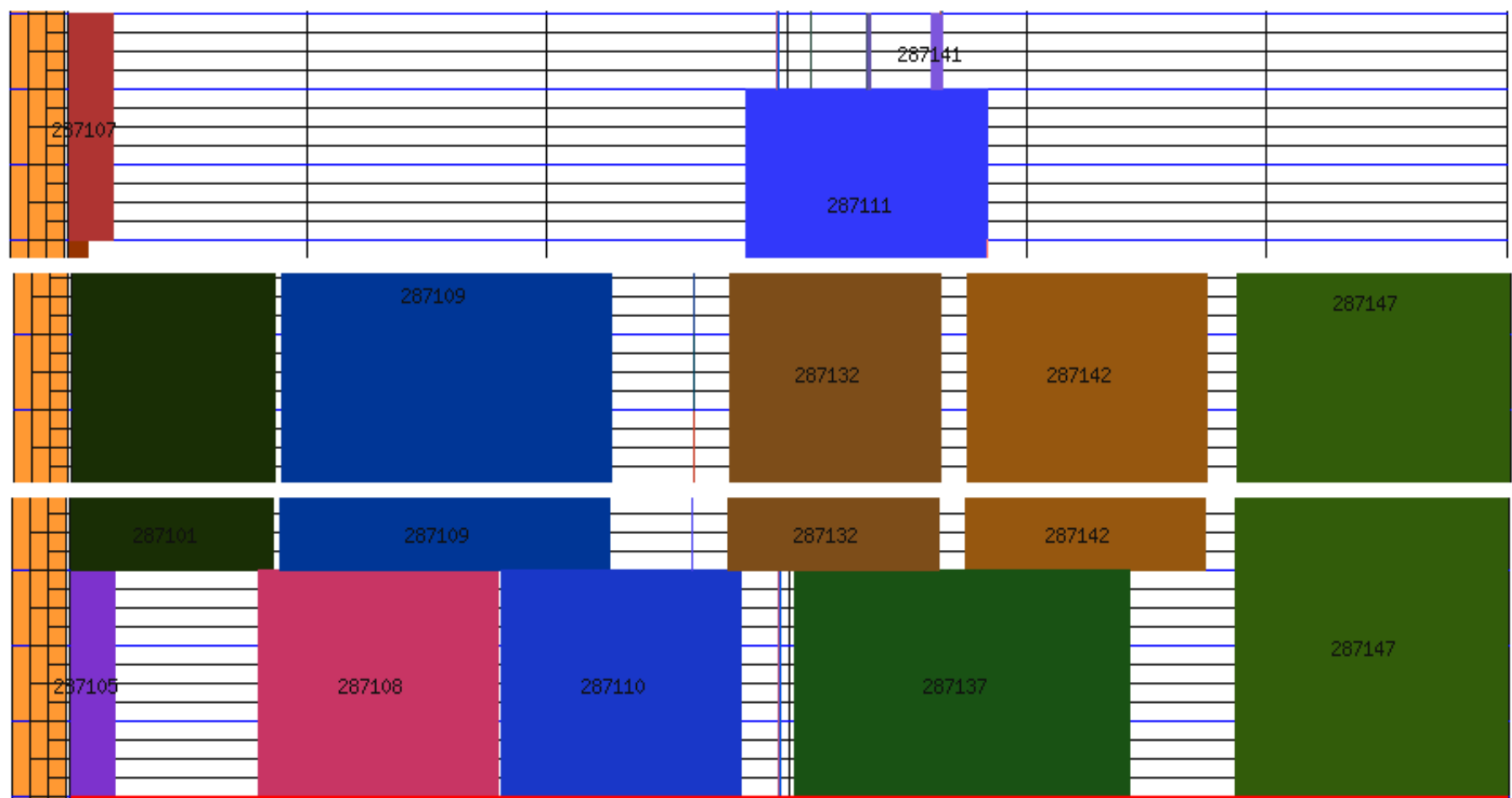




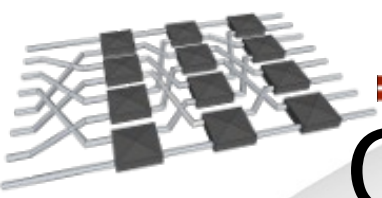
Temporal Access Pattern

- Applications present interleaved phases of computation & I/O
 - Idleness during **processing phases**
 - High I/O rate during **Input** or **Output phases**
- Constant rate of I/O
 - Application may have a long input or output phase
 - I/O done in the background while application executes

Concurrent Execution of Applications

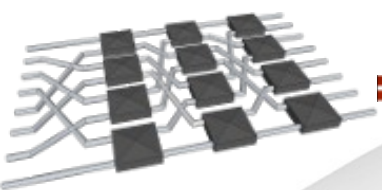


Toulouse Site of Grid5000, Aug 2nd - 5th



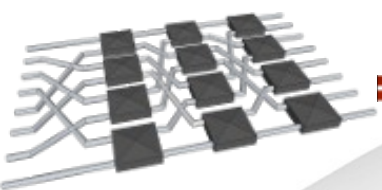
Concurrent Execution of Applications

- Concurrent Access to a Shared Storage System
 - Shared I/O & Network Bandwidth to the servers
- Any combination of access patterns
- *How will the bandwidth of one application behave due to the access patterns of others ?*



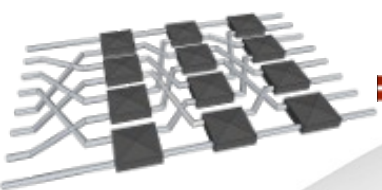
Methodology

- MPI-IO-Test v21
 - Writes/Reads **Objects** of a given **Size**.
 - Optionally, waits a specified interval between each operation
- Pastel Cluster, Toulouse site of Grid5000
- 24 NFS Clients
- dNFSp with 6 Servers – Each acting as meta-server and IOD

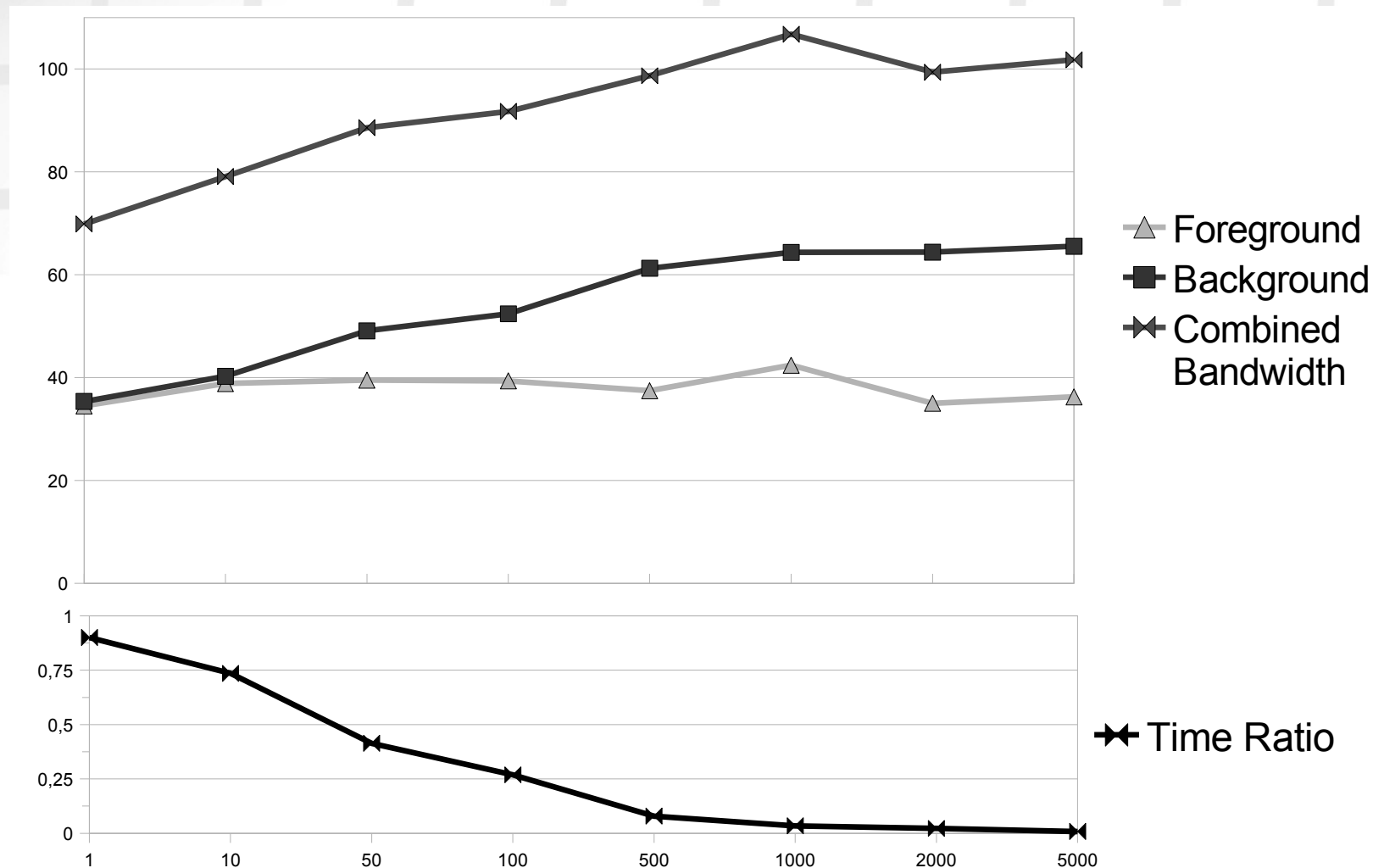


Methodology

- Clients divided in 2 sets:
 - **Background:** Waits a specified interval between each I/O operation
 - **Foreground:** No interval in between operations
- Concurrent execution during 3 minutes.
 - Write as many objects as it can
- Objects sizes of **128KB**, **2MB** and **4MB**



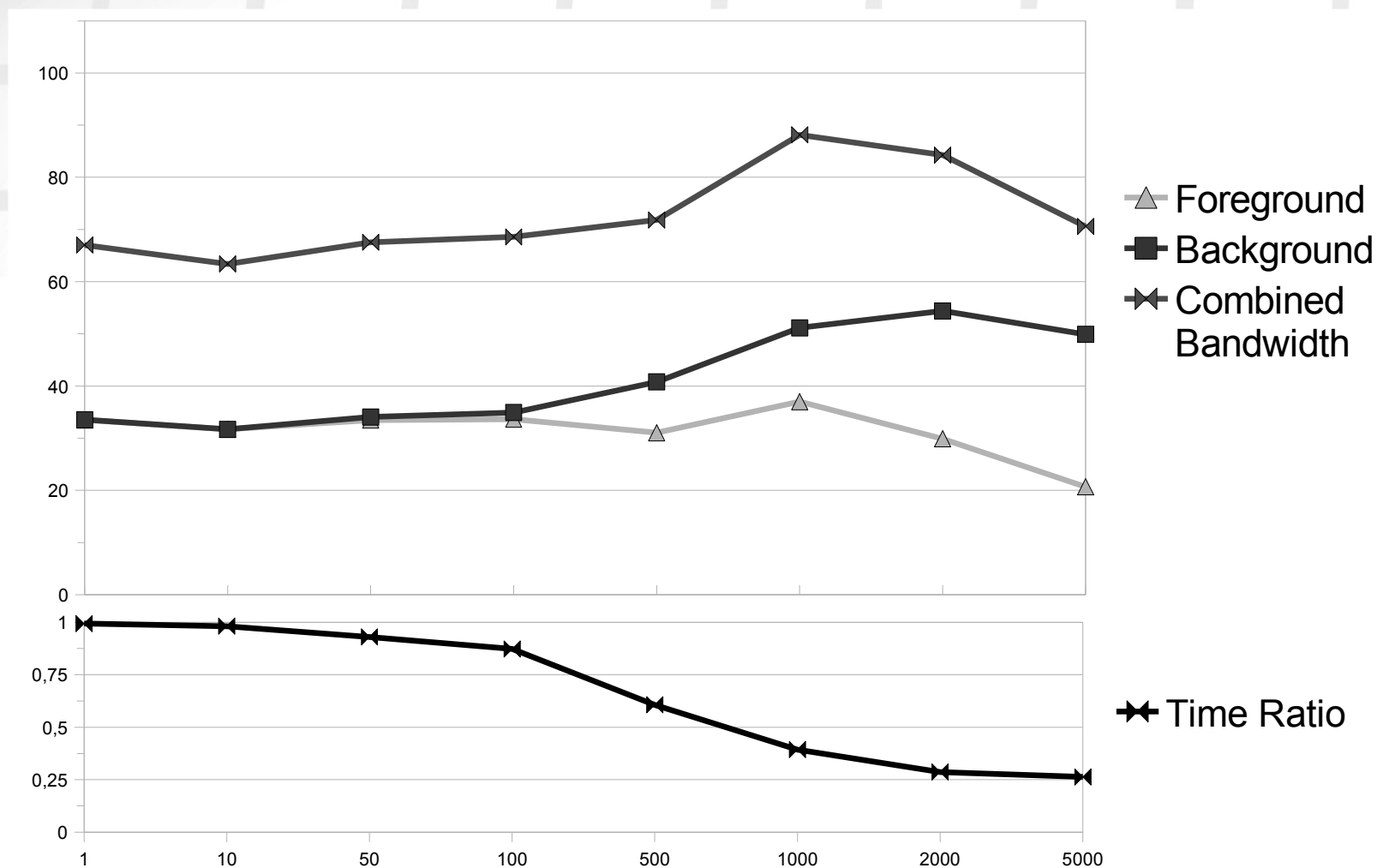
Results – 128KB Objects

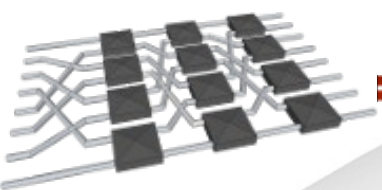


Results – 128KB Objects

- Divergence started with interval of 50ms
- Foreground bandwidth ranges from 34MB/s to 36MB/s (42MB/s peak at interval of 1s)
- Background ranged from 35MB/s to 65MB/s
- Combined bandwidth reached 100MB/s

Results – 2MB Objects

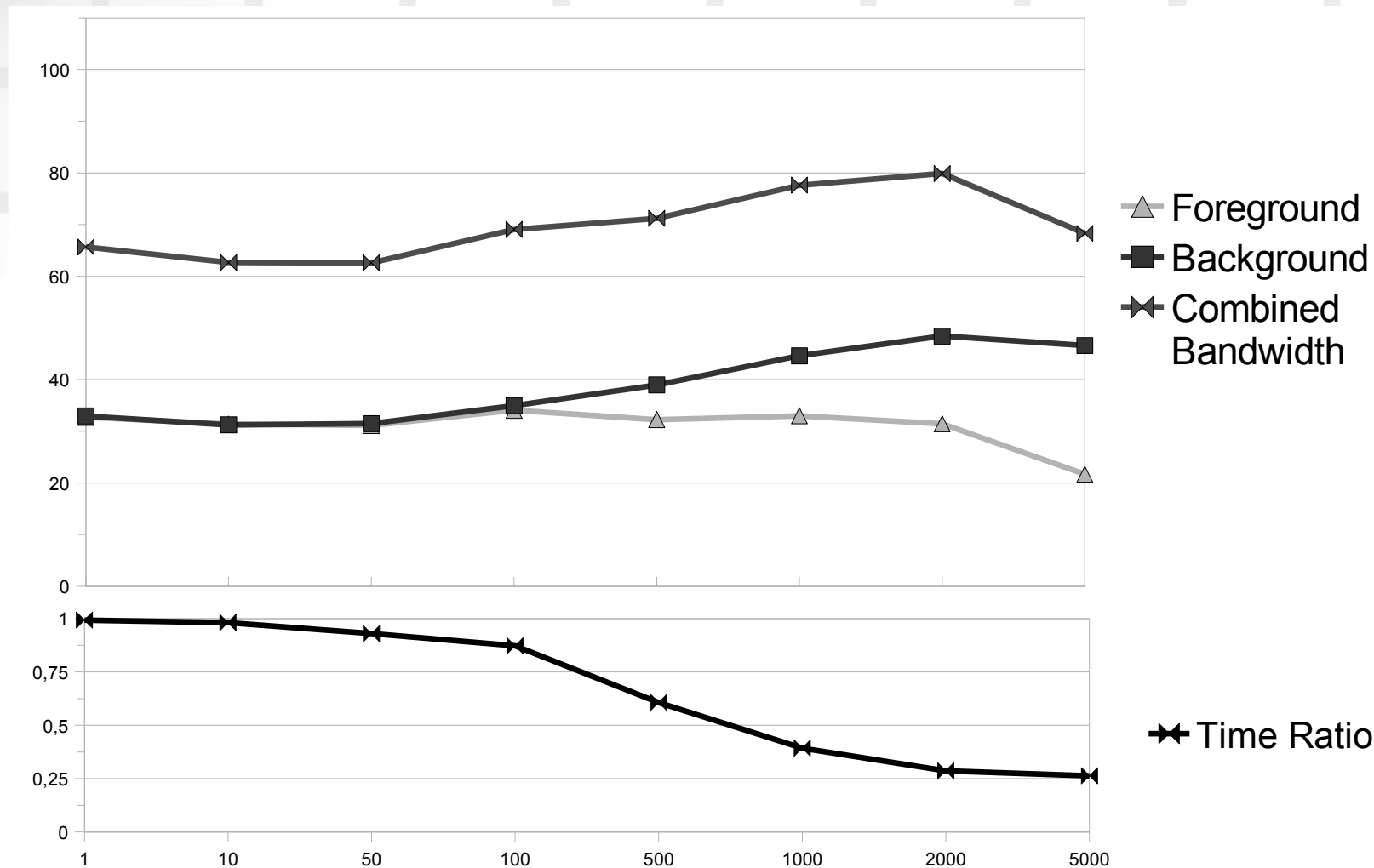


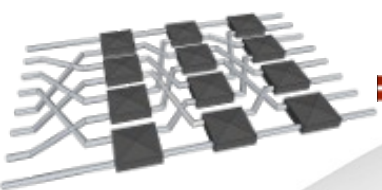


Results – 2MB Objects

- Foreground ranged from 33MB/s to 20MB/s, peaking 37MB/s at 1s
- Background ranged from 33MB/s to 49MB/s
- dNFSp whole utilization was below expected with longer writes: 88MB/s peak

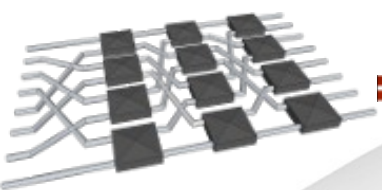
Results – 4MB Objects





Results – 4MB Objects

- Foreground ranged from 33MB/s to 21MB/s
- Background ranged from 33MB/s to 46MB/s
- Combined Bandwidth peaked ~80MB/s with 2s interval



Conclusions

- Longer writes have shown **worse performance** than short ones
 - Likely an effect of delayed-write client politics.
- With bigger object sizes, foreground instance has decreased performance when interval grows
- For the tested object sizes, a **time ratio** of 0,75 seems to be the limit where the intervals are influential to other executions

[Interaction of Access Patterns on the dNFSp File System]

Rodrigo Virote Kassick

