Leistungsanalyse
Hochleistungs-Ein-/Ausgabe

Michael Kuhn

Wissenschaftliches Rechnen
Fachbereich Informatik
Universität Hamburg

2015-06-22

E/A-Schichten

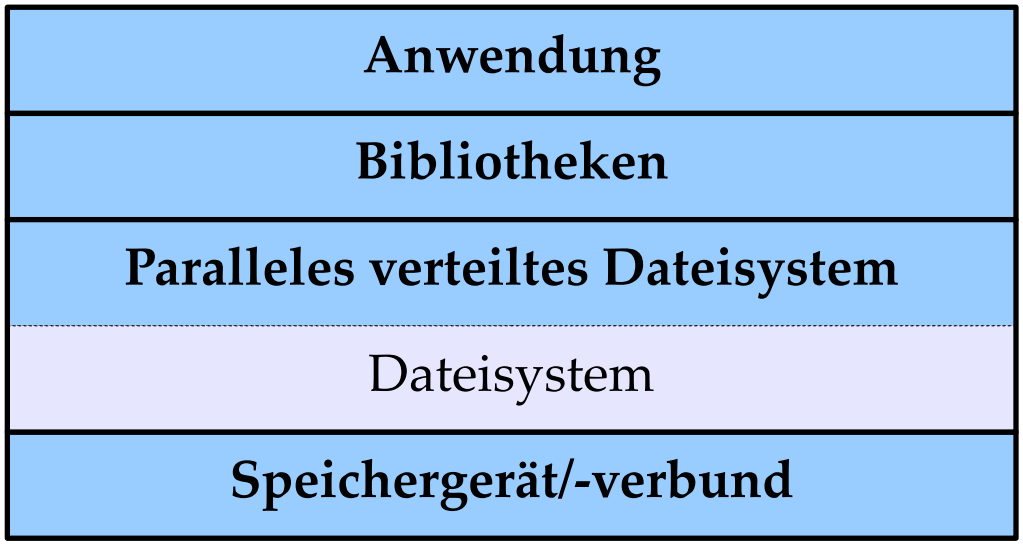


Abbildung: E/A-Schichten

Einleitung

- Leistungsanalyse ist ein schwieriges Thema
  - Software und Hardware immer komplexer
  - Viele Schichten und Interaktionen
- Analyse besteht aus zwei Teilen
  - Messung und Bewertung
- Messung erfasst aktuelle Leistungsfähigkeit
  - Richtig messen ist auch schwierig
- Bewertung erlaubt Vergleich mit möglicher Leistung
  - Insbesondere bei Beschaffungen etc.

Benchmarks

- Unmengen an Benchmarks verfügbar
  - Für alle möglichen Aspekte
  - CPU, RAM, Netzwerk etc.
- Für E/A ebenso
  - IOzone, Bonnie, Bonnie++, PostMark, b_eff_io, FLASH I/O und viele mehr
  - Vorgestellt: fio, IOR und mdtest
- Üblicherweise Erfassung eines bestimmten Zugriffsmusters
  - Daher viele verschiedene Benchmarks für viele verschiedene Anwendungsfälle

fio [2]

- Flexibler E/A-Tester
  - Autor ist Jens Axboe, Maintainer des Block-Layers
  - Unter anderem verantwortlich für die CFQ-, NOOP- und Deadline-Scheduler sowie blktrace und `splice()`
- Beliebige Workloads
  - Statt vieler spezieller Tests
- Unterstützung durch Job-Dateien
  - Gemeinsame und job-spezifische Parameter
  - Alles auch über Kommandozeile steuerbar

fio…

- Operationstypen
  - Lesen/schreiben/gemischt, sequentiell/zufällig
- Blockgröße
  - Sowohl einzelne Werte als auch Bereiche
- Datengröße
  - Wie viele Daten insgesamt zugegriffen werden sollen
- E/A-Engine
  - Synchron/asynchron, Memory Mapping, null
- Warteschlangentiefe
  - Für asynchrone E/A-Engines

fio…

- Buffered, direct, fsync
  - Zur Messung bzw. Umgehung des Caches
- Datei- und Threadanzahl
  - Parallele Workloads
- Sperren
  - Keine, exklusiv, nicht-exklusives Lesen
- Preallokation mit `fallocate`
  - Messung der Blockallokation
- Optimierung mit `fadvise`
  - Hinweise bezüglich Zugriffsmuster

fio…

- Ausrichtung
  - Beispielsweise zur Ausrichtung der E/A an Seitengrenzen
- Komprimierbar- und Deduplizierbarkeit
  - Aktuelle SSDs komprimieren die Daten transparent
- Durchsatzlimit
  - Nützlich um Hintergrundlast zu erzeugen
- Verifikation
  - Überprüfung ob die gelieferten Daten den vorher geschriebenen entsprechen

fio…

```
1 [global]
2 rw=randread
3 size=128m
4
5 [job1]
6
7 [job2]
```

Listing 1: fio-Beispiel [2]

- Zufälliges Lesen aus einer 128 MiB großen Datei
- Zwei Prozesse `job1` und `job2`
  - Dateinamen werden automatisch generiert

fio…

```
1 [random-writers]
2 ioengine=libaio
3 iodepth=4
4 rw=randwrite
5 bs=32k
6 direct=0
7 size=64m
8 numjobs=4
```

Listing 2: fio-Beispiel [2]

- Asynchrone E/A mit Tiefe 4
- Vier Prozesse, die zufällig schreiben
  - Eigene, 64 MiB große Dateien

fio…

- Unterstützt auch Trace-Replay
  - Erlaubt Ausführung aufgezeichneter E/A-Muster
  - Eigentliche Anwendung nicht notwendig
  - Vergleich mit anderen Systemen
- blktrace und eigenes Format
  - `write_iolog` und `read_iolog`
  - ASCII-Format und leicht selbst erstellbar

IOR [1]

- Für parallele verteilte Dateisysteme Zugriff von mehreren Knoten notwendig
  - fio erlaubt nur mehrere Prozesse auf einem Knoten
  - IOR nutzt MPI für parallelen Zugriff
- Unterstützung für mehrere Backends
  - POSIX, MPI-IO, HDF5 und Parallel-NetCDF
- Einige unterschiedliche E/A-Modi
  - Gemeinsame oder prozess-lokale Dateien
  - Prozessumordnung zur Umgehung von Cachingproblemen
    - Client X schreibt, Client X+n liest

IOR...

```
IOR START
  api=MPIIO
  testFile=/path/to/file
  repetitions=3
  readFile=0
  writeFile=1
  filePerProc=0
  keepFile=0
  segmentCount=10
  blockSize=1g
  transferSize=1m
RUN
IOR STOP
```

- Schreiben in gemeinsame Datei mit drei Wiederholungen

IOR… [3]

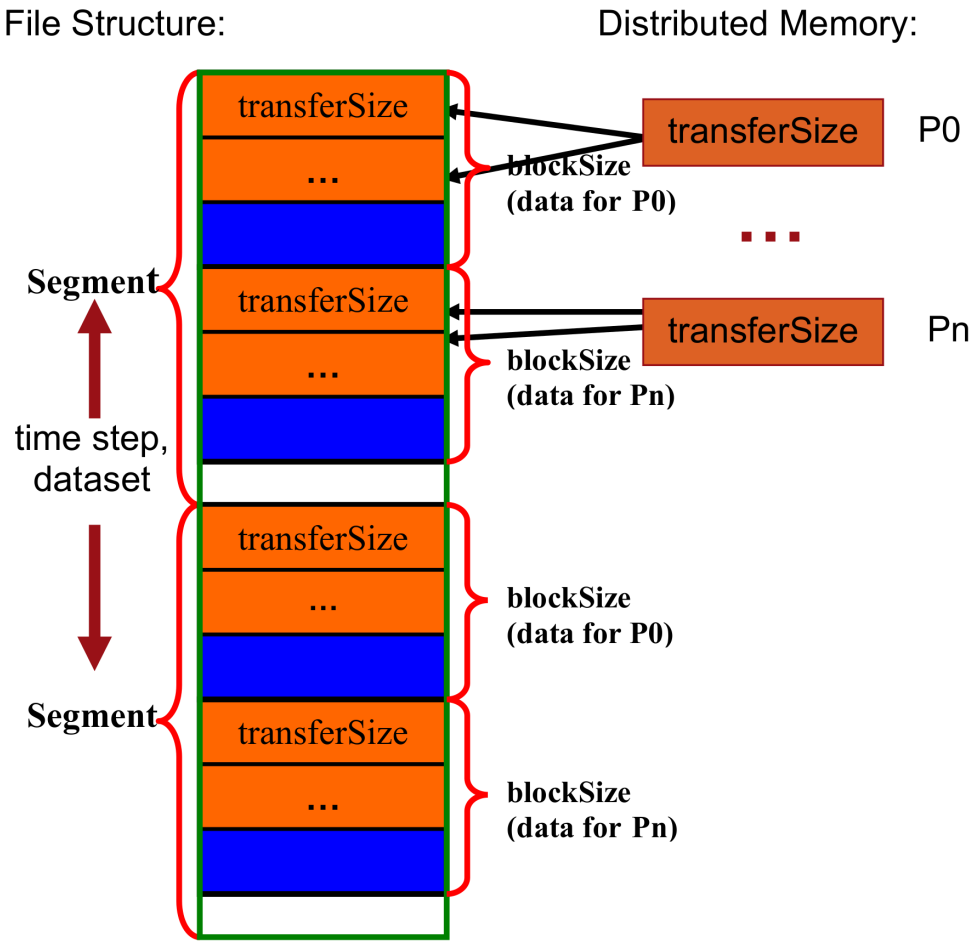


Fig. 1. The design of the IOR benchmark for shared file type. Blocks are stored in separate files for the 1-file-per-processor mode of operation.

mdtest

- Datendurchsatz durch Benchmarks gut abgedeckt
  - Metadaten aber auch ein wichtiger Faktor
- mdtest nutzt MPI für parallelen Metadatenzugriff
  - Operationen werden mit POSIX-Schnittstelle durchgeführt
  - MPI-IO bietet nicht ausreichend Funktionalität
- Aufgeteilt in drei Phasen
  - Erstellen, Status abrufen und Löschen
- Arbeitet mit hierarchischer Struktur

mdtest…

- Unterstützt ähnliche Funktionalitäten wie IOR
  - Prozessumordnung für Cache-Umgehung
  - Mehrere Wiederholungen
- Optional können Daten in Dateien geschrieben werden
  - Bei Bedarf auch synchronisiert werden
- Unterstützung für mehrere Wurzelverzeichnisse
  - Beispielsweise um mehrere Metadatenserver zu nutzen

Vergleich

- Vielzahl von Benchmarks für unterschiedlichste Anwendungsfälle
  - Sowohl Daten als auch Metadaten
- Ergebnisse allerdings nicht vergleichbar
  - Unterschiedliche Muster
  - Unterschiedliche Berechnung
  - Unterschiedliches Verhalten
- Unklarheiten bezüglich Ergebnissen
  - MB vs. MiB

Werkzeuge

- Benchmarks erlauben Erfassung der momentanen Leistung
  - Kein Aufschluss über Ursachen etc.
- Analyse und Optimierung erfordern Werkzeuge
  - Einblick in innere Vorgänge notwendig
  - Üblicherweise mit Hilfe von Tracing (VampirTrace)
- Manchmal abstrakte Betrachtung ausreichend
  - Grobe Übersicht über E/A-Verhalten (Darshan)

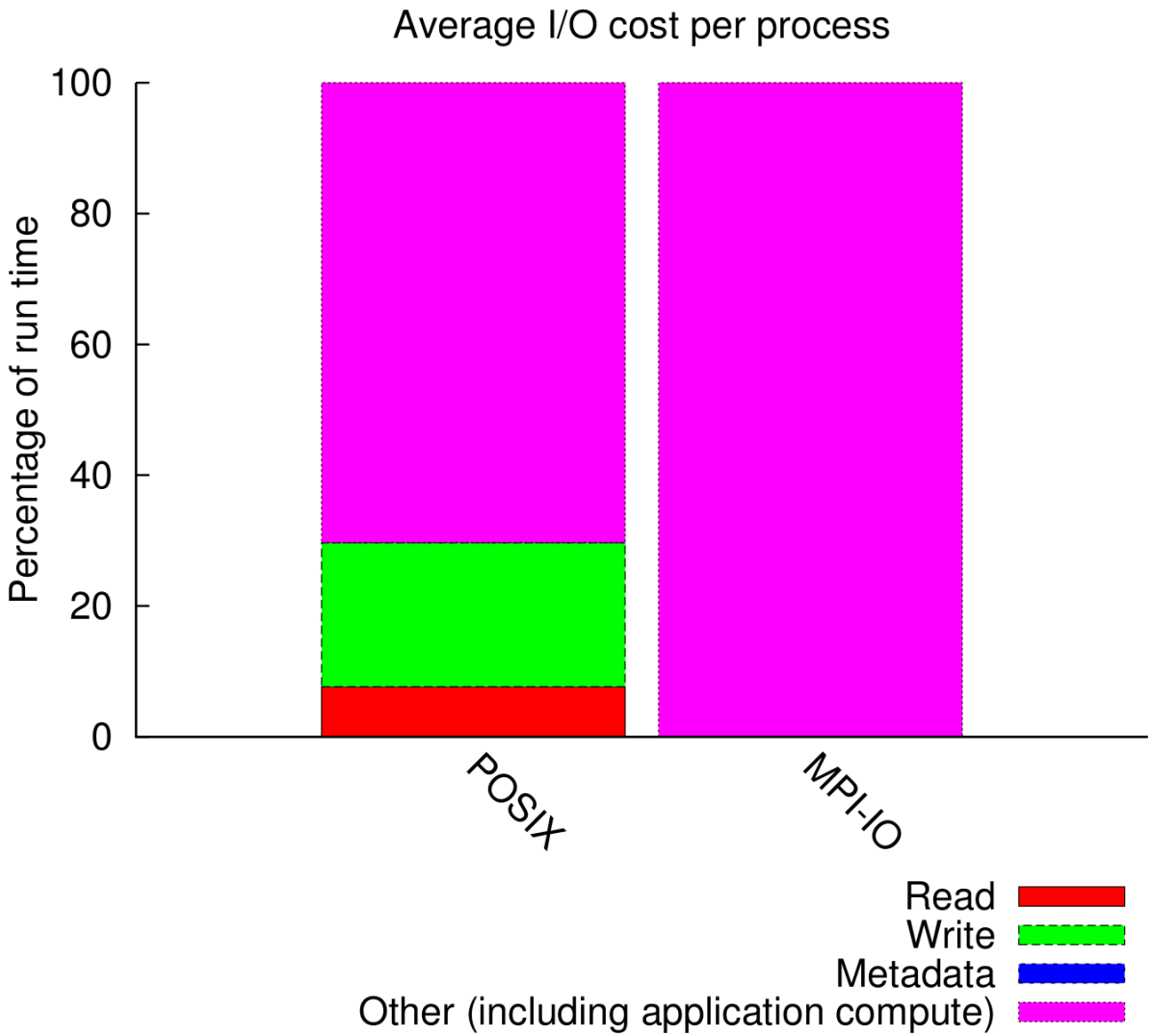
Darshan

- Werkzeug zur E/A-Charakterisierung
  - Sanskrit für „Sicht“ oder „Vision“
- Möglichst genaues Bild der Anwendungs-E/A
  - Einschließlich Eigenschaften wie der E/A-Muster
  - Dabei aber minimaler Overhead
- Geeignet für dauerhaften Einsatz
  - Wurde mit über 750.000 Kernen getestet
- Sehr gute Unterstützung für MPICH
  - Argonne National Laboratory
  - Gruppe um OrangeFS, MPICH und ROMIO

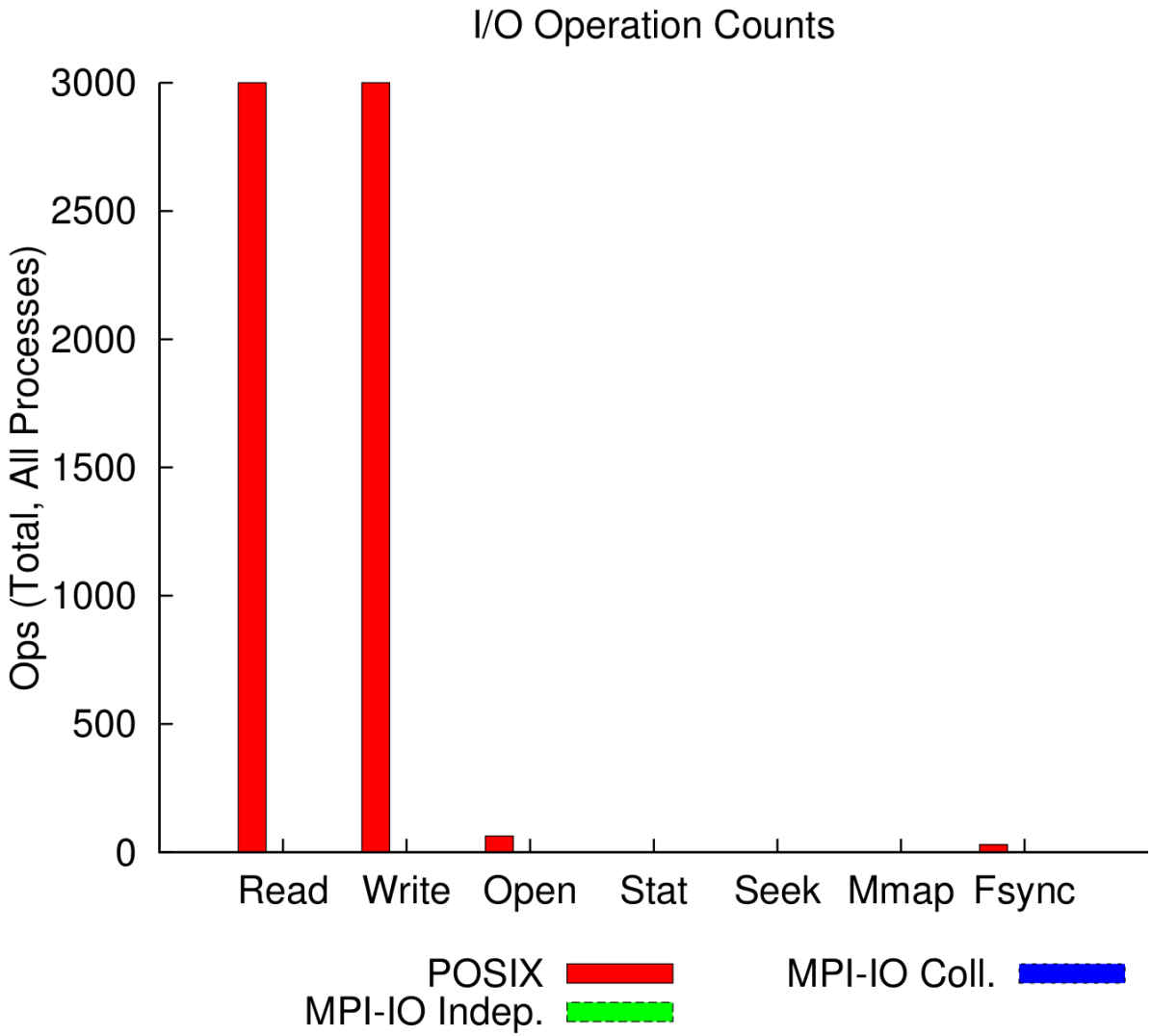
Darshan…

- Besteht aus zwei Teilen
  - Runtime und Werkzeuge
- Runtime erfasst Anwendungs-E/A
  - Spezifisch für eine MPI-Implementierung
  - Außerdem Optionen für Batch-Scheduler und gemeinsames Protokollverzeichnis
  - Compiler-Wrapper und Preload-Bibliothek `libdarshan.so`
- Werkzeuge analysieren aufgezeichnete Protokolle
  - Befehle: `darshan-job-summary.pl`, `darshan-parser` etc.

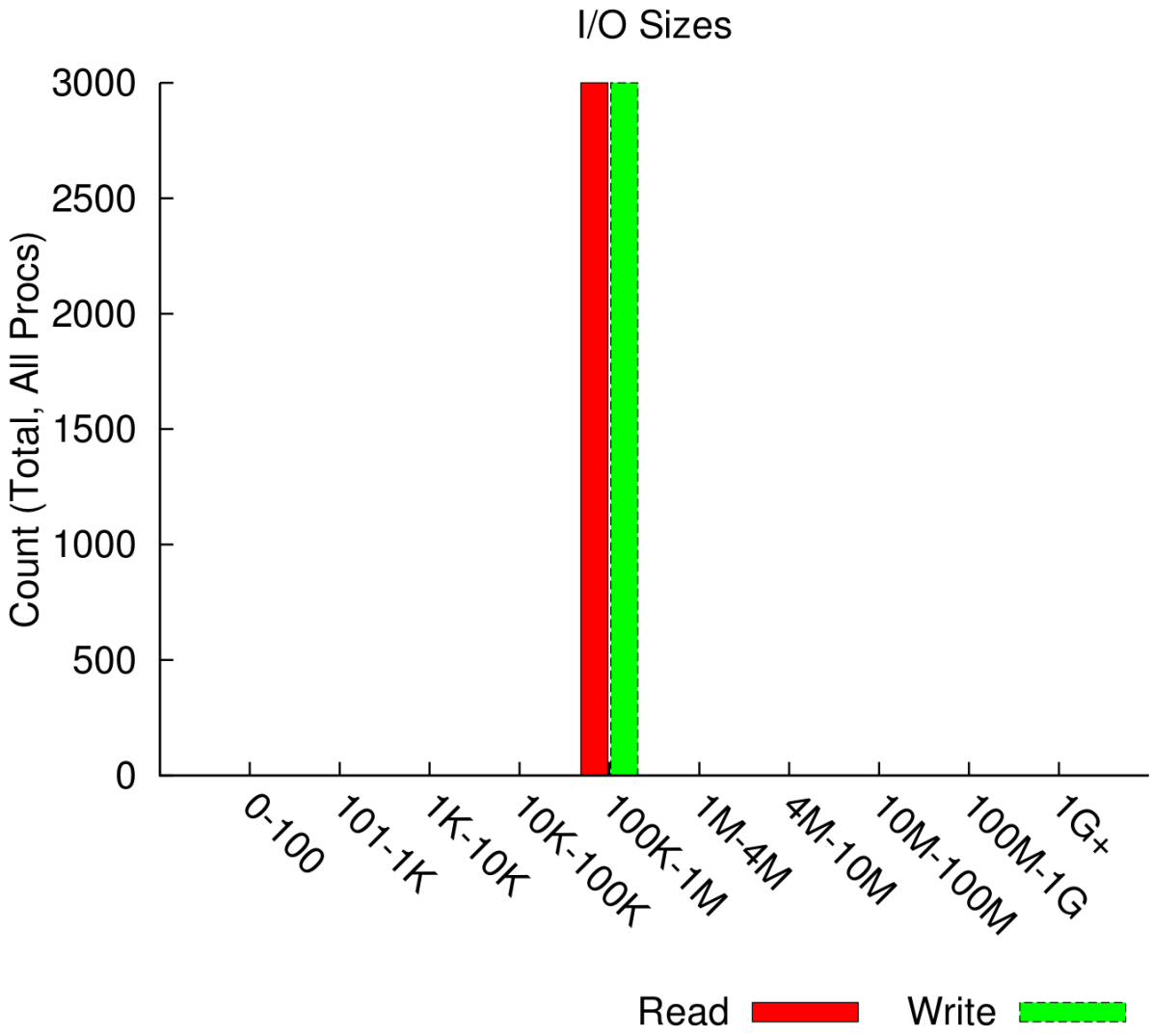
Darshan…


Average I/O cost per process
Percentage of run time
100
80
60
40
20
0
POSIX
MPI-IO
Read
Write
Metadata
Other (including application compute)

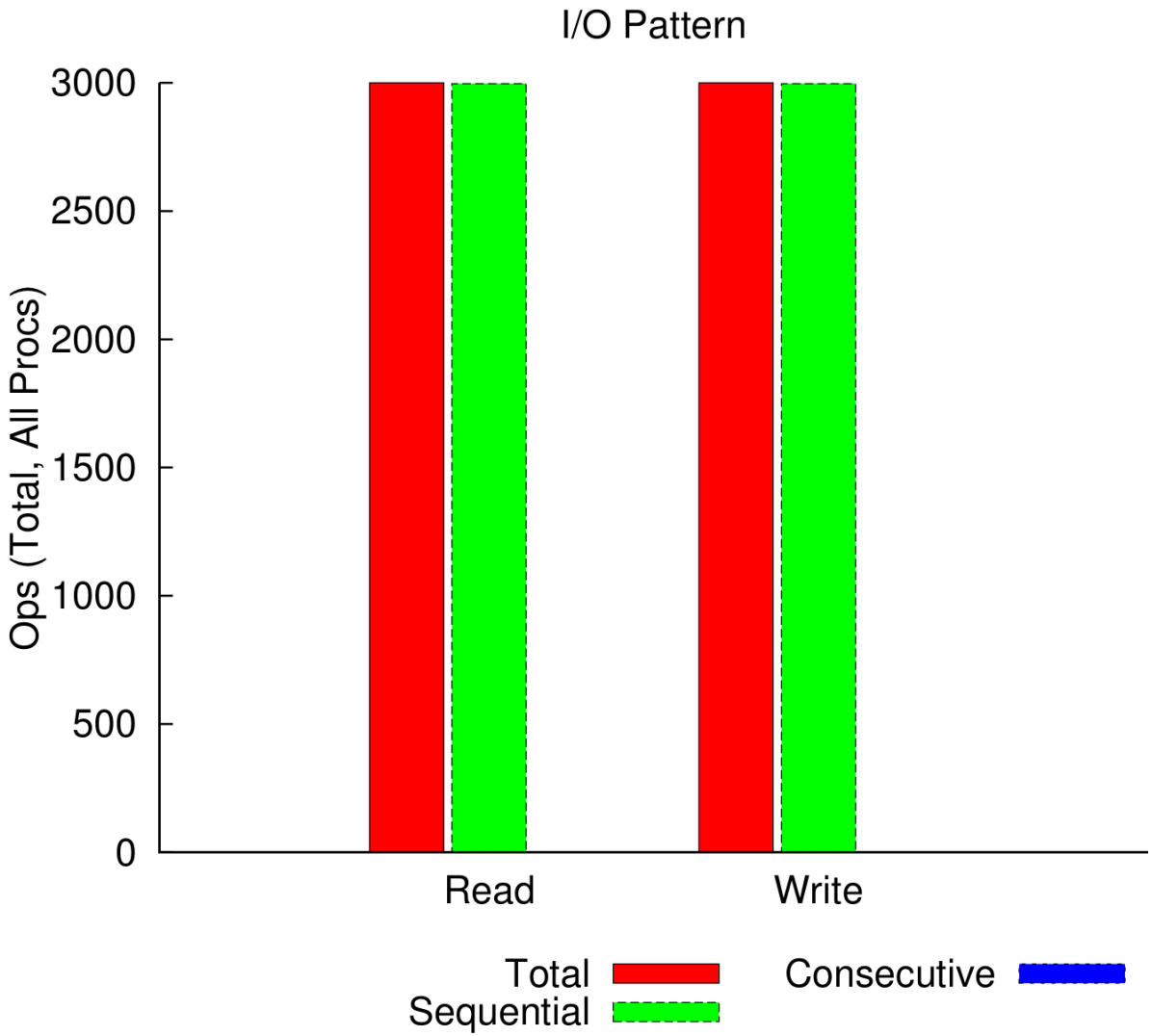
Darshan…


I/O Operation Counts
Ops (Total, All Processes)
3000
2500
2000
1500
1000
500
0
Read
Write
Open
Stat
Seek
Mmap
Fsync
POSIX
MPI-IO Indep.
MPI-IO Coll.

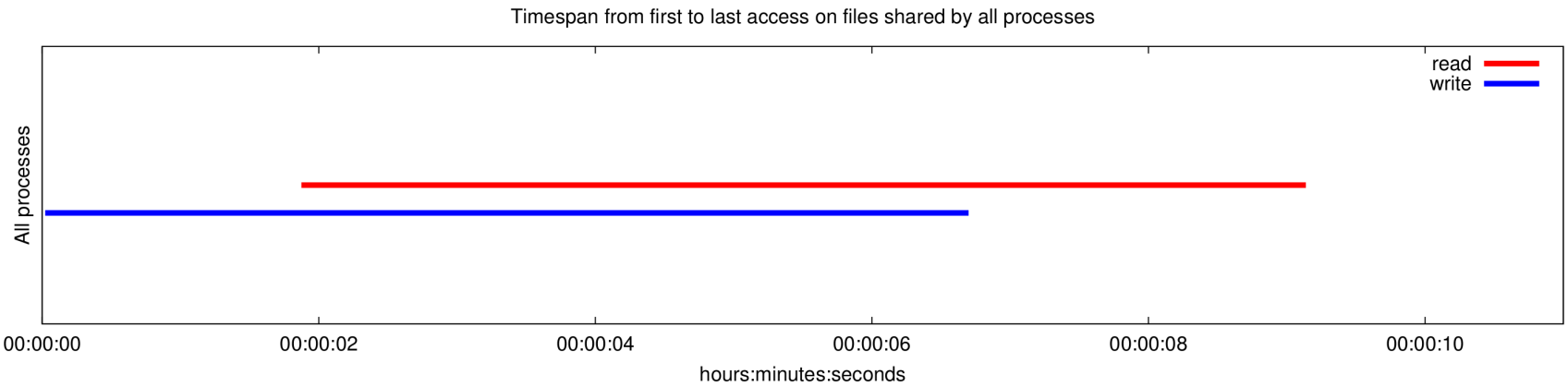
Darshan...


I/O Sizes
Count (Total, All Procs)
3000
2500
2000
1500
1000
500
0
0-100
101-1K
1K-10K
10K-100K
100K-1M
1M-4M
4M-10M
10M-100M
100M-1G
1G+
Read
Write

Darshan…


I/O Pattern
Ops (Total, All Procs)
3000
2500
2000
1500
1000
500
0
Read
Write
Total
Sequential
Consecutive

Darshan…

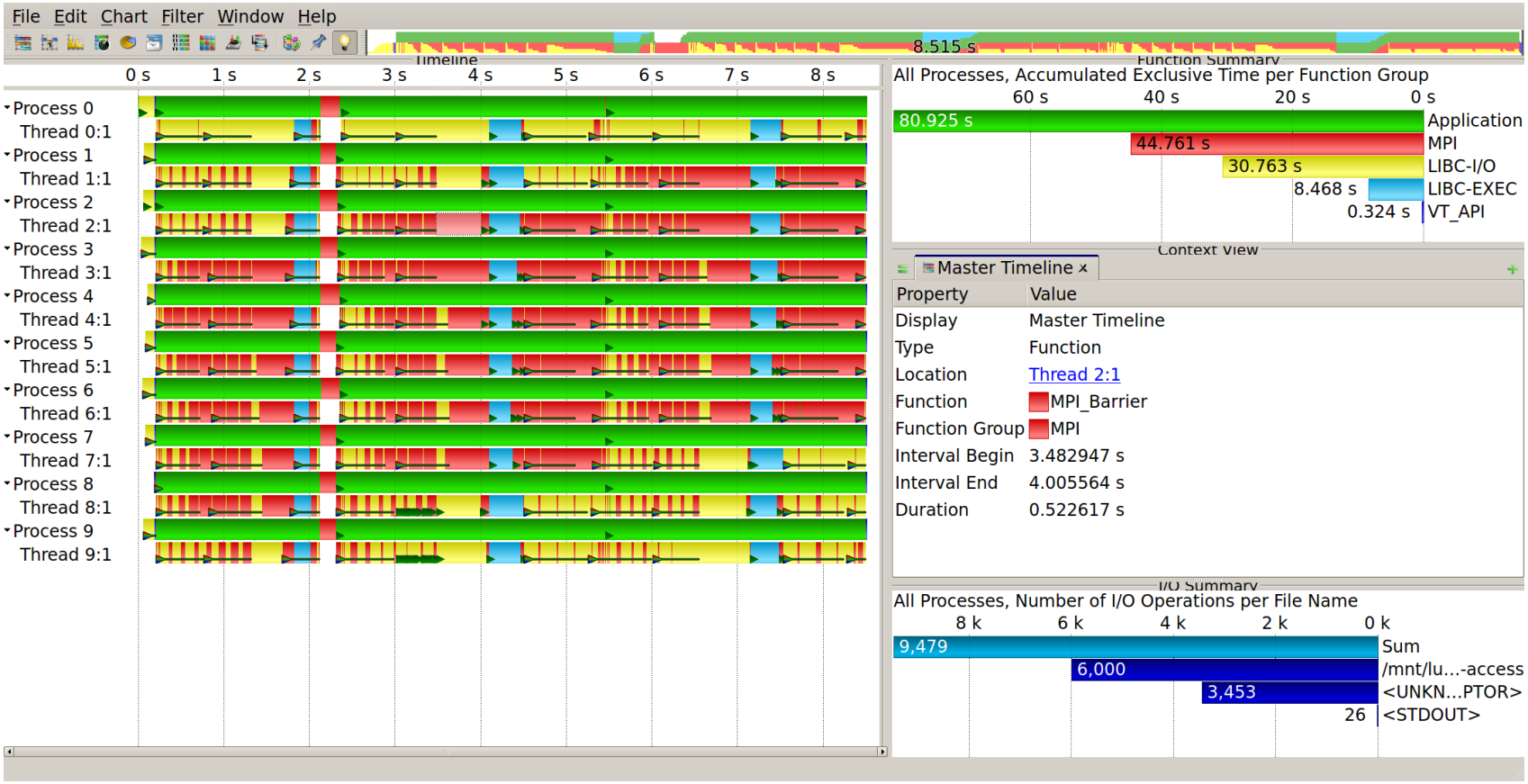


- Grobe Übersichten der Kosten von E/A
  - Bezüglich Aufrufanzahl, Zugriffsgröße und -muster
- Erlaubt Abschätzung ob Optimierung notwendig ist

VampirTrace

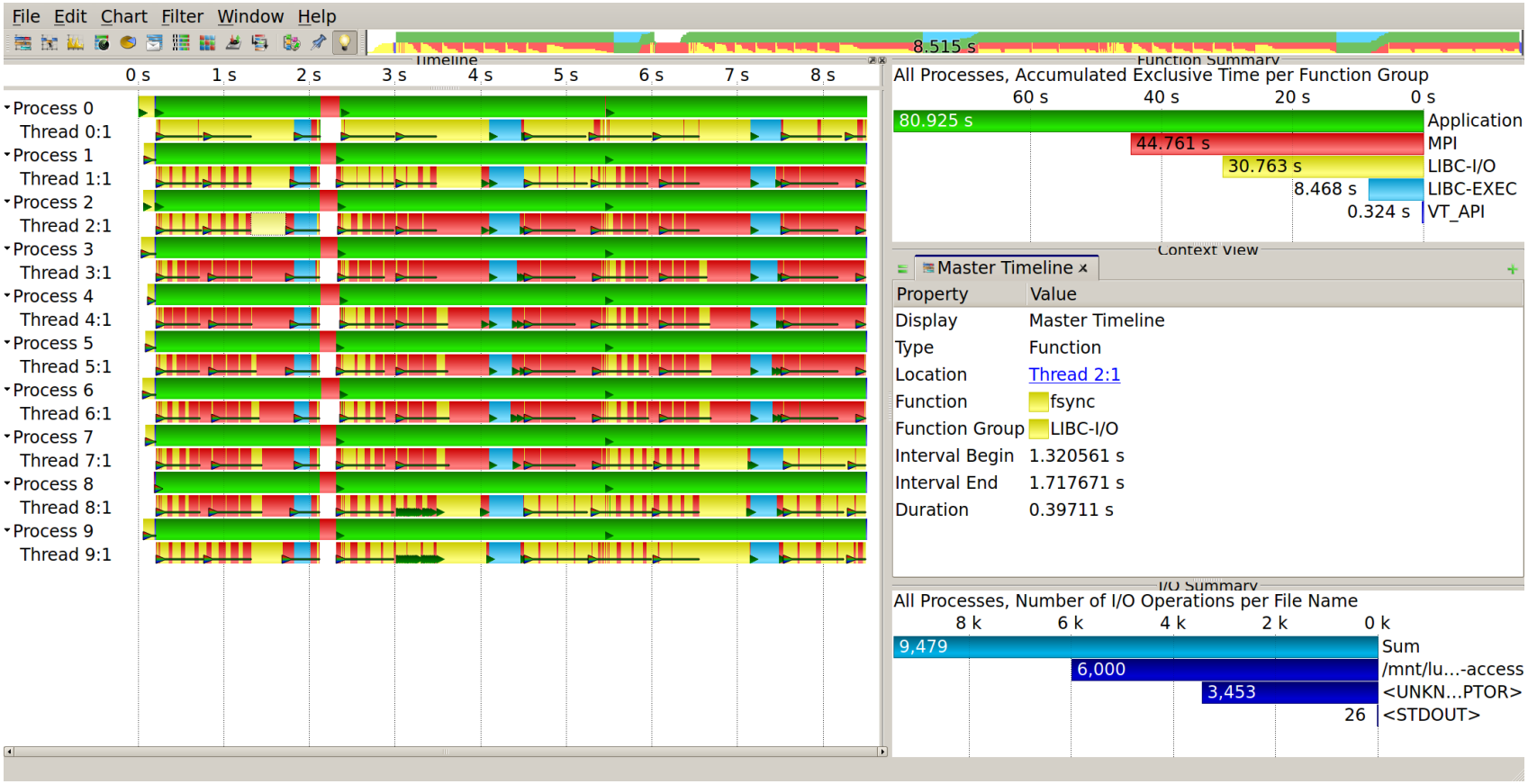
- VampirTrace zeichnet Spurdaten auf
  - VampirTrace ist Open Source
  - Nachfolger: Score-P
  - Spezifisch für eine MPI-Implementierung
  - Compiler-Wrapper `vtcc`
- Vampir zeigt Spurdaten an
  - Vampir ist kommerziell
  - Evaluationslizenzen verfügbar
- Spurdaten sind deutlich größer als Darshan-Protokolle
  - Im getesteten Fall mehr als Faktor 100

VampirTrace…



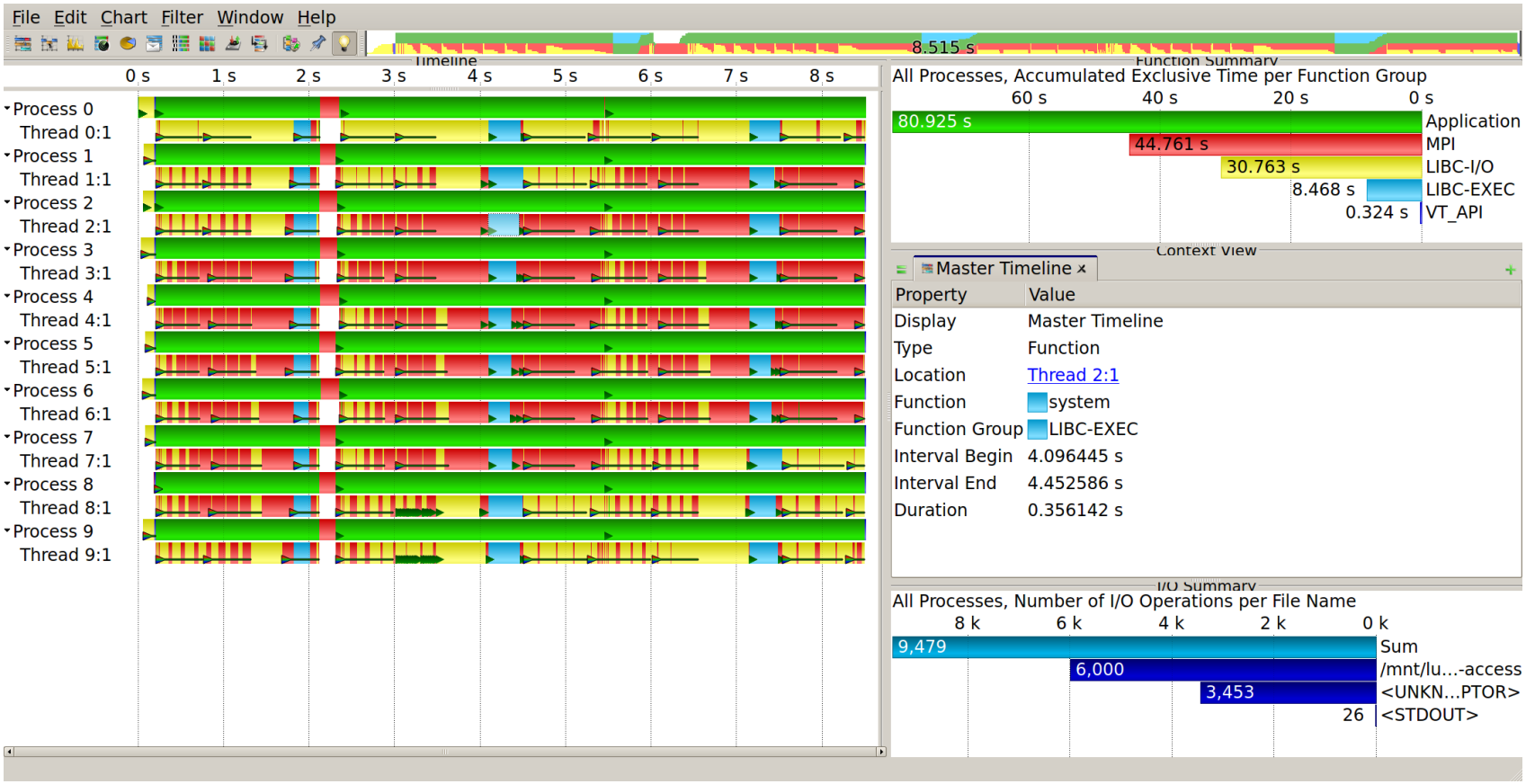
- Sehr ungleiche E/A-Zeiten, dadurch lange Barries

VampirTrace…



- Synchronisieren dauert teilweise sehr lange

VampirTrace…

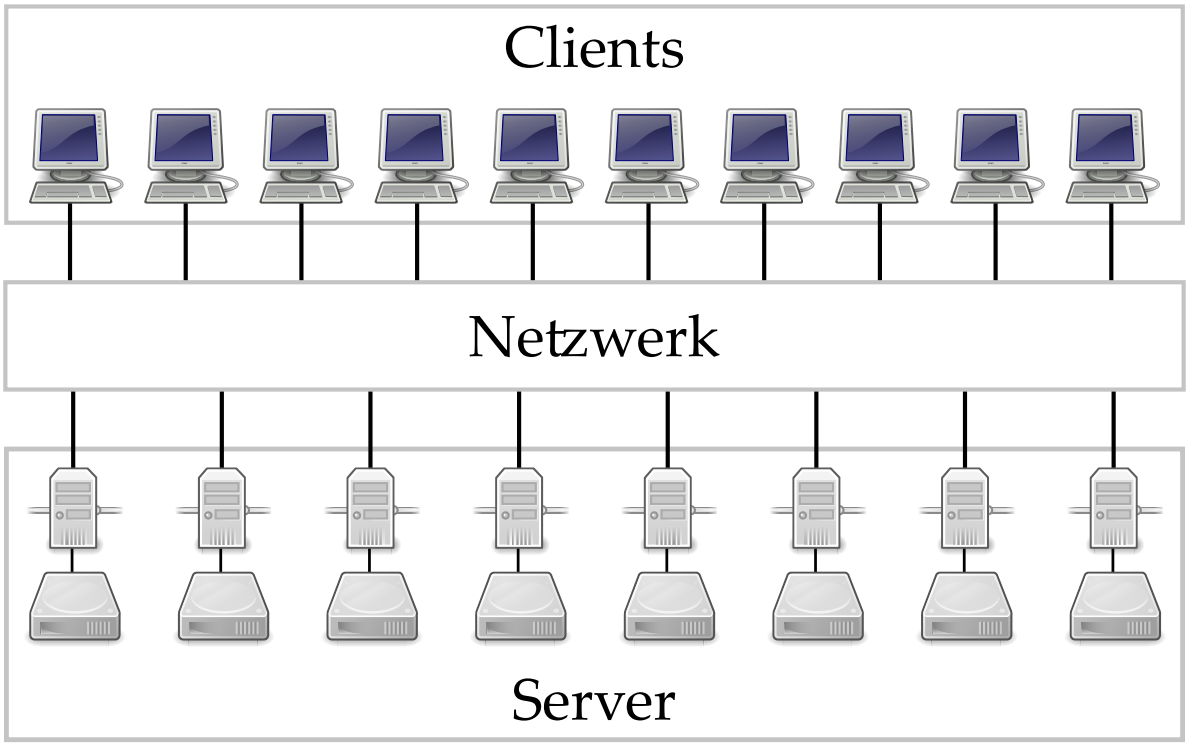


- Leeren des Caches auch sichtbar

Einleitung

- Bewertung der Leistung durch Modellierung der theoretisch möglichen Leistung
- Dazu sind einige Informationen notwendig
  - Involvierte Komponenten
  - Leistungscharakteristika der Komponenten
- Zusätzliche Leistungsmessungen der Komponenten
  - Dazu wieder andere Werkzeuge

Einleitung…



- Clients: IOPS, RAM, Bus zum Netzwerk
- Netzwerk: Durchsatz und Latenz
- Server: Durchsatz und IOPS der Speichergeräte, Bus

Client

- Anzahl der E/A-Operationen pro Sekunde
  - Kontextwechsel könnten Beschränkung sein
- Durchsatz und Latenz des Hauptspeichers
  - Üblicherweise kein Problem
- Abschätzung mit Hilfe von `tmpfs` und `fio`

```
1 $ mkdir /tmp/fs
2 $ mount -t tmpfs tmpfs /tmp/fs
3 $ ...
4 $ umount /tmp/fs
```

Client…

```
1 $ fio --name=switch --filename=/tmp/fs/foo --rw=write --bs=1
    ↪ --size=1g --runtime=60 [--numjobs=n]
2 $ vmstat 1
3 $ fio --name=bw --filename=/tmp/fs/foo --rw=write --bs=1m
    ↪ --size=$size --runtime=60
```

- Messung auf `west`
- ≈ 1.000.000 IOPS
- ≈ 300.000 Kontextwechsel
- ≈ 3 GiB/s Durchsatz
  - Hauptspeicher vermutlich höher, da Overhead durch `tmpfs`

Netzwerk

- Deutliche Unterschiede je nach Netzwerktechnologie
  - InfiniBand vs. Ethernet
- Durchsatz des Netzwerks
  - Muss höher sein als Leistung der Speichergeräte
- Anzahl der Pakete pro Sekunde
  - Beschränkung bei vielen kleinen Nachrichten
- Messung mit Hilfe von `ping` und `iperf`

Netzwerk…

```
1 $ ping -c 60 $host
2 $ iperf --server --port $port
3 $ iperf --client $host --port $port
```

- Messung zwischen `west` und `sandy`
- Round Trip Time ≈ 0,230 ms
  - Entspricht ≈ 4.350 Nachrichten pro Sekunde
  - Geringfügig niedriger auf zweiter Netzwerkschnittstelle
- Durchsatz ≈ 112 MiB/s

Speichergerät

- Deutliche Unterschiede je nach Speichertechnologie
  - HDD vs. SSD
- Durchsatz der Speichergeräte
  - Niedriger als Netzwerkdurchsatz für maximale Ausnutzung
  - Höher als Netzwerkdurchsatz für Leistungsreserven
- IOPS wichtig für Metadatenoperationen
  - Auch für zufällige Datenzugriffe
- Speicherbus kann auch beschränkender Faktor sein
  - Teilweise immer noch SATA 2.0 (300 MB/s)

Speichergerät…

```
1 $ fio --name=iops --filename=/dev/sd? --direct=1 --rw=randread
    ↪ --bs=4k --size=$size --runtime=60
2 $ fio --name=bw --filename=/dev/sd? --direct=1 --rw=read
    ↪ --bs=1m --size=$size --runtime=60
```

- HDDs
  - IOPS ≈ 60–80
  - Durchsatz ≈ 120 MiB/s
- SSDs
  - IOPS ≈ 15.000
    - Ausreißer mit ≈ 5.500
  - Durchsatz ≈ 270 MiB/s

Beispiel

- Messung des Datendurchsatzes
  - Lustre und OrangeFS
- Unterschiedliche Blockgrößen
  - 1 MiB und 64 KiB
  - Entspricht den Streifenbreiten von Lustre und OrangeFS
- Bei 4.350 Nachrichten pro Sekunde
  - 4,25 GiB/s (1 MiB) bzw. 272 MiB/s (64 KiB)
- Flaschenhals ist Netzwerkdurchsatz
  - Maximal 1.120 MiB/s

Beispiel…

Dateisystem	Blockgröße	1 PPN	6 PPN	12 PPN
Lustre	1 MiB	640 MiB/s	105 MiB/s	110 MiB/s
OrangeFS	1 MiB	160 MiB/s	390 MiB/s	430 MiB/s
OrangeFS	64 KiB	250 MiB/s	115 MiB/s	180 MiB/s

Tabelle: Schreiben

Dateisystem	Blockgröße	1 PPN	6 PPN	12 PPN
Lustre	1 MiB	1.095 MiB/s	735 MiB/s	875 MiB/s
OrangeFS	1 MiB	130 MiB/s	265 MiB/s	430 MiB/s
OrangeFS	64 KiB	505 MiB/s	140 MiB/s	195 MiB/s

Tabelle: Lesen

- Massiver Leistungseinbruch bei Lustre
- Streifenbreite als Blockgröße teilweise vorteilhaft

Beispiel…

Dateisystem	Blockgröße	1 PPN	4 PPN	8 PPN
Lustre	$^{1}/_{\text{PPN}}$ MiB	640 MiB/s	620 MiB/s	605 MiB/s
OrangeFS	$^{64}/_{\text{PPN}}$ KiB	250 MiB/s	280 MiB/s	210 MiB/s

Tabelle: Schreiben

Dateisystem	Blockgröße	1 PPN	4 PPN	8 PPN
Lustre	$^{1}/_{\text{PPN}}$ MiB	1.095 MiB/s	1.800 MiB/s	525 MiB/s
OrangeFS	$^{64}/_{\text{PPN}}$ KiB	505 MiB/s	655 MiB/s	455 MiB/s

Tabelle: Lesen

- Bessere Leistung in beiden Fällen
- Anomalie beim Lesen von Lustre

Zusammenfassung

- Benchmarks erlauben die Erfassung der aktuellen Leistung
  - Viele unterschiedliche Benchmarks für viele unterschiedliche Anwendungsfälle
- Werkzeuge sind notwendig zur eingehenden Analyse
  - Sowohl für Grobüberblick als auch zur Detailanalyse
- Messung sagt nichts über mögliche Leistungsfähigkeit aus
  - Dafür ist eine Bewertung der Leistung notwendig
  - Häufig durch Leistungsmodellierung

Zusammenfassung…

- Grobe Modellierung ist häufig schon ausreichend
  - Lässt sich bei Bedarf verfeinern
- Unvorhersehbares Verhalten macht Analyse schwierig
  - Siehe beispielsweise Leseleistung bei Lustre
- Analyse der konkreten Implementierung notwendig
  - Optimierung setzt viele Detailkenntnisse voraus

Quellen I

[1] CHAOS Development Team. IOR - Parallel filesystem I/O benchmark. https://github.com/chaos/ior.

[2] Jens Axboe. fio - Flexible IO Tester. http://git.kernel.dk/?p=fio.git;a=summary.

[3] Hongzhang Shan and John Shalf. Using IOR to Analyze the I/O Performance for HPC Platforms. In *In: Cray User Group Conference (CUG'07*, 2007.