JÜLICH
FORSCHUNGSZENTRUM

Characterizing Parallel I/O Behaviour Based on Server-Side I/O Counters

SC16 - BoF Analyzing Parallel I/O

SC16 BoF - Analyzing Parallel I/O, November 15, 2016 | **S. El Sayed**[JSC] M. Bolten[Kas] D. Pleiter[JSC] |
[JSC] Jülich Supercomputing Centre, s.elsayed@fz-juelich.de

[Kas] Institut für Mathmatik, Universität Kassels

Motivation

Why analyse I/O?

- I/O to compute imbalance
- Applications I/O requirements are increasing

Solution: Emerging I/O architectures

- Hierarchical storage
- Active storage

Key Point

Impact of emerging I/O architectures requires understanding I/O load characteristics on current high-end HPC systems

Methodology
Performance Counters

- Assuming an I/O sub-system that periodically (Δt) logs 6 values (for an extended time):
  - Data read [Bytes]
  - Number of read operations [IOP]
  - Number of file open operations
  - Data written [Bytes]
  - Number of write operations [IOP]
  - Number of file close operations
- Collect job (Application run during I/O logging) information
  - Start time, end time and I/O servers used
- Link performance counters to job

Characterisation Criteria

- Category 1: Aggregate performance numbers
  1. Total amount of data read/written
  2. Total number of IOPs
  3. Read/Write bandwidth
  4. Read/Write IOPS
  5. Total number of files created
  6. I/O intensity

- Category 3: Parallel I/O
  1. **Parallel I/O intensity**
  2. I/O operation concurrency
  3. Parallel I/O distribution
  4. Same file access concurrency

- Category 2: I/O pattern analysis
  1. Distribution of request sizes
  2. Percentage of small I/O
  3. Request size: Variable vs fixed
  4. Percentage of I/O type
  5. Dominating I/O operation type
  6. Task-local vs shared
  7. Spatial access patterns
  8. Temporal distribution of I/O
  9. **Burstiness parameter**
  10. Access pattern repetitive behaviour
  11. Dominating I/O operation repetitiveness

Characterisation Criteria

Basic Quantities

c is a threshold parameter with $c \geq 0$

$\delta(s, t, \Delta t)$ Helper quantity $= 1$ if more than c Bytes are moved on server s

$H(t, \Delta t)$ Helper quantity $= 1$ if more than c Bytes are moved on any server.

$$H(t, \Delta t) = \begin{cases} 1 & \delta(s, t, \Delta t) > 0 \quad \text{for any server } s\,, \\ 0 & \text{otherwise} \end{cases}$$

Characterisation Criteria

Burstiness

Considering:

l_{IO} Average number of consecutive intervals Δt with $H = 1$

l_{noIO} Average number of consecutive intervals Δt with $H = 0$

2.9 Burstiness parameter

$$\rho = \begin{cases} 1 - \tanh(l_{\mathrm{IO}}/l_{\mathrm{noIO}}) & \text{if } l_{\mathrm{noIO}} > 0\,, \\ 0 & \text{otherwise} \end{cases}$$

$tanh$ bounds burstiness parameter to the interval [0,1].

Key Point

As l_{IO} increases ρ tends to 0, while when l_{noIO} increases ρ tends to 1, where $0 \leq \rho \leq 1$.

Characterisation Criteria

Parallel I/O intensity

$|S|$ Number of I/O servers used by the job.

$\pi(t, \Delta t)$ Fraction of servers involved in I/O during $[t, \Delta t]$

3.1 Parallel I/O intensity

$$\Pi = \frac{\sum_i \pi(t_s + i\Delta t, \Delta t)}{\sum_i \delta(t_s + i\Delta t, \Delta t)}$$

Normalised:

$$P = \frac{|S|\,\Pi - 1}{|S| - 1}$$

$P = 1$ when I/O $> c$ all I/O servers are involved

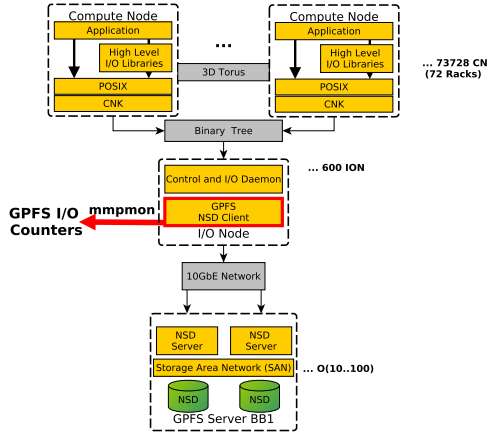
$P = 0$ when I/O $> c$ only one I/O server is involved

Selected Results

I/O sub-system background

- JUGENE (72 racks of BlueGene/P)
- I/O sub-system uses GPFS
- Performance counters logged on the 600 I/O nodes with $\Delta t = 120s$ for approximately 19 months
- Analysed 0.17 million jobs that ran over 1 hour

Counter	Description
_br_	Bytes read
_bw_	Bytes written
rdc	Read requests
_wc_	Write requests



Selected Results

I/O intensity, burstiness & Parallel I/O intensity

- 80% of analysed jobs are equal or below these values

Threshold c	0 Byte read	128 KiByte read	1 MiByte read
Burstiness (ρ)	0.99	0.99	1.0
Parallel I/O intensity (P)	0.91	0.88	0.84

Threshold c	0 Byte write	128 KiByte write	1 MiByte write
Burstiness (ρ)	0.0	1.0	1.0
Parallel I/O intensity (P)	1.0	0.28	0.27

Future Work

- GPFS performance counters monitoring has been enabled on all large scale-systems at Jülich Supercomputing centre
- Monitoring data has been integrated into LLview
- We plan to apply the characterisation metrics to collected data and integrate these into LLview

Thanks.
QUESTIONS?

JÜLICH
FORSCHUNGSZENTRUM

BACKUP!

Linking Performance Counters and I/O Criteria

$D_r(l, s, t)$ Number of read operations of length l Bytes arriving at server s during **[**t_s**,**t**]**

$D_w(l, s, t)$ Number of write operations of length l Bytes arriving at server s during **[**t_s**,**t**]**

$\delta(s, t, \Delta t)$ Helper quantity with value 1 if more than c Bytes are moved

For GPFS:

Counter	Description	Observable
`_br_`	Bytes read	$\sum_l l\, D_r(l, s, t)$
`_bw_`	Bytes written	$\sum_l l\, D_w(l, s, t)$
`_rdc_`	Read requests	$\sum_l D_r(l, s, t)$
`_wc_`	Write requests	$\sum_l D_w(l, s, t)$

$$\delta_r(s, t, \Delta t) = \begin{cases} 1 & \text{if } \sum_l l\,[D_r(l, s, t + \Delta t) - D_r(l, s, t)] > c\,, \\ 0 & \text{otherwise} \end{cases}$$

where $c \geq 0$ is a threshold parameter.