Knowing your Enemy:  
Addressing the I/O bottleneck by Profiling

Andra Hugo  
Jean-Thomas Acquaviva  
DDN
DDN Advanced Technical Center

- DDN, world leader in HPC storage
  - present in 70% of the TOP500
  - 650 persons WW, ⅔ in engineering

- R&D centers
  - France, Meudon Emerging tech and Software Defined Storage
    → 25+ R&D engineers
  - Japan
  - US East Coast
  - US West Coast
  - India, Pune
Problem Statement – Diversity of stacks

- From HPC to HPDA
- Disruptive innovation of storage systems
  - SSD, NVMe
  - 1000x less latency

- Difficult to understand
  - Scale out analysis

- Difficult to optimize
  - Isolate bottlenecks
Problem Statement – Diversity of loads

Application specific optimizations are not enough!
I/O profile impact on performance: meets IO500

As expected, the log-structure approach in IME means very little performance is lost from ior_easy_write to ior_hard_write.

BeeGFS has an interesting spread. Some great results at the bottom! But some bad ones at the top. The next slide tries to figure out why...

Easy:
- Writes of 1Mo, sequential
- 1 process/file

Hard:
- Writes of 47Ko, random
- 1 shared file
IDIOM: Integrated Device I/O Monitor

► **Main target:**
  - Accelerate & partially automate I/O optimizations
  - Insure performance portability on new storage systems

► **FUI 25: Fond Unique Interministériel**
  - Industry oriented, 1 call per year
  - Tightly coupled to “pole de compétitivité” by Groupes Thématiques

► **Value proposition**
  - Monitor and Characterize IO workloads HPC & HPDA
    - Identify hotspot
    - Propose optimization
    - Identify most suitable storage backend
  - Monitoring and Tracing tool
    - To be deployed from laptop to data center
    - Capture applications I/O with an overhead < 3%

► **700 KE of funding**
IDIOM partners gathering forces

- **DDN Storage**
  - I/O application tracing
- **Criteo**
  - Multi file systems applications
- **Qarnot computing**
  - Distributed systems
- **QuasarDB**
  - Time Series databases for IOT
- **CEA-DAM**
  - Deployment in production systems

- **Telecom SudParis**
  - I/O x86, ARM tracing
- **Université de Bretagne Occidentale**
  - I/O kernel tracing
- **INRIA Grenoble**
  - I/O aware task scheduling
Towards a standard I/O profiling tool

- DDN Dio-Pro
  - Application: Tracing tool for IO characterization

- SupTelecom ParisSud EzTrace
  - Application / SystemTracing tool support x86 / ARM

- UBO VFSMon, FuncMon, and iotracer,
  - Kernel: Low level from laptop to large system

Build a chain of tools exploitable in an industrial context
Main challenges in complex systems

- **Parallelism:**
  - Synchronization in a distributed system
  - Aggregation of parallel execution traces

- **Depth**
  - Multi-level traces

- **Coverage**
  - Two application stacks: HPC and HPDA

- **Execution overhead**

- **Diversity of deployment environments**

- **Define I/O patterns**
  - Automatic learning
ID.IO.M working plan

- User & kernel land information gathering
- Application characterization & I/O system dimensioning
- Infrastructure management for deployment
- API definition for the applications (including visualization)
- ID.IO.M + HPC batch-scheduler -> I/O aware scheduler
- HPC application analysis
- Distributed system validation
- File system impact analysis on I/O
- Smart building application validation
Conclusion: IDIOM’s main objectives

► Address the data deluge in a pragmatic way
► I/O characterization of HPC & HPDA applications
► Deployment on different systems: from laptop to datacenters

► Collect data to understand
► Accelerate & partially automate I/O optimizations
► Insure performance portability on new storage systems

► Kick-Off last October… Still much to do …
Thank you!