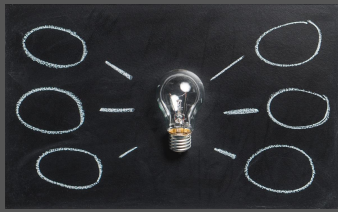
Department of Computer Science

HPS ACES

University of Reading

Strategic Planning Agenda



Limitless Storage
Limitless Possibilities

https://hps.vi4io.org

Julian M. Kunkel

ACES Strategy Meeting

2019-03-25

LIMITLESS **POTENTIAL** | LIMITLESS **OPPORTUNITIES** | LIMITLESS **IMPACT**

Outline

University of Reading

Department Activities

University of Reading

Teaching

- CS1PC – Programming in C
- CS3DP – Distributed and Parallel Computing

Administrative

- SID: Strategic infrastructure development for the department
- Museum: C37 (history of computer science, ongoing)

Research Interest

University of Reading

High-performance storage for HPC

- Efficient I/O
  - Performance analysis methods, tools and benchmarks
  - Optimizing parallel file systems and middleware
  - Modeling of performance and costs
  - Tuning of I/O: Prescribing settings
  - Management of workflows
- Data reduction: compression library, algorithms, methods
- Interfaces: towards domain-specific solutions and novel interfaces

Other research interests

- Application of big data analytics (e.g., for transportation)
- Domain-specific languages (for Icosahedral climate models)
- Cost-efficiency for data centers in general and for “produced” science

Research

University of Reading

Involvement in currently running projects

- AIMES (ends in Q3/2019) – DFG funded
  Advanced Computation and I/O Methods for Earth-System Simulations
- PeCoH (ends in Q1/2020) – DFG funded
  Performance Conscious HPC
- Cooperation with Bull (ends Q4/2020) – industry funded
  I/O Analysis at DKRZ
- ESiWACE (ends Q3/2019) – H2020 project
  Centre of Excellence in Simulation of Weather and Climate in Europe
  Moved some money to Reading for PDRA (to start ASAP)

Projects in the Queue

- Advanced Storage Monitoring (2PM PDRA) / Cooperation with DDN
- ESiWACE2 (3Y PDRA) follow up of ESiWACE

Co-Supervised PhDs

University of Reading

University of Hamburg

- Anastasiia Novikova: Compression of Climate/Weather Data
- Frank Gadban: Understanding the Convergence of HPC and Cloud Computing
- Nabeeh Jumah: Language Extensibility and Configurability to Support Stencil Code Development
- Eugen Betke: Machine Learning of I/O Behavior
- ... (focus on relevant ones)

University of Reading

- Ekene Ozioko: Coordination Scheme for Human Driven and Autonomous Vehicle Traffic Intersection using traffic light with intersection control unit
- Haifa Alsultan (start 04/2019?): Big Data Processing for Climate Science

Support Activities

University of Reading

- Regularly attended conferences (others infrequently)
  - ISC-HPC (June)
  - Supercomputing (November)
- Community building
  - Bootstrapped: The Virtual Institute for I/O
    https://www.vi4io.org
  - Supporting: European Open File System (EOFS)
    https://www.eofs.eu/
  - Organizing: Various I/O workshops / Birds-of-a-feather sessions
- Awareness: co-created the IO-500 list
  http://io-500.org
- Beyond teaching:
  - Online teaching platform for C-Programming (ICP project)
  - Establishing a **HPC certification program**
    https://hpc-certification.org

Long-Term Software-Development

University of Reading

- IOR/MDTest: Benchmarks for HPC I/O
- ESDM: Earth-System Data Middleware (ESiWACE)
- SCIL: Scientific Compression Library
- Various small benchmarks/tools

Outline

University of Reading

Near-Term Goals

University of Reading

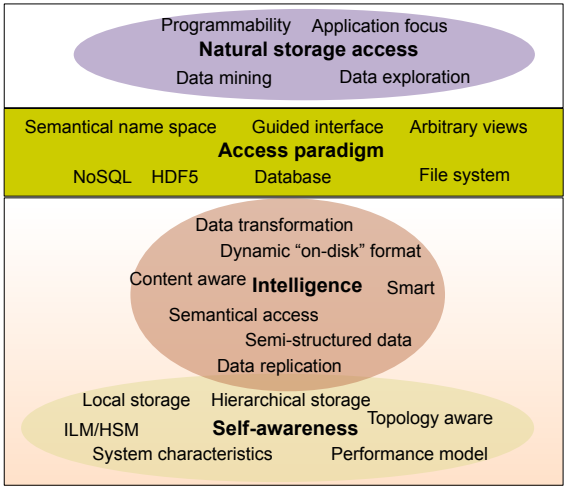
- Get the PostDoc (Luciana Pedro) started for ESiWACE and deliver...
- Integrate compression into long-term archives (JASMIN and others)
- Make I/O Middleware ready for "production" runs and enabling machine learning
- In-memory storage and compute (cooperation with KOVE/Argonne)
- First certificates of the HPC Certification Program go live
  - Tailor HPC Certification Program for the need of NWP community
- Comparison of DSLs for Weather Climate on the Shallow Water Equation GDDML/GridTools/PsyClone (joint paper)
- (Rework the modules for CS1PR|PC / CS3DP)
- (Opening the museum)

Outline

University of Reading

Personal Vision: Towards Intelligent Storage Systems and Interfaces

University of Reading



- Abstract data interfaces
- Enhanced data management
- Integrate compute/storage engine
- Flexible views on data
- Smart storage
  - Self-aware systems
  - AI optimized placement
- Across sites

ESDM is just the Beginning: Next Generation Interfaces

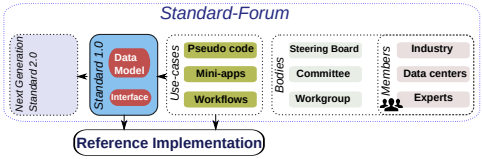
University of Reading

Towards a new I/O stack considering:

- User metadata and workflows as first-class citizens
- Smart hardware and software components
- Liquid-Computing: Smart-placement of computing
  - Utilizing arbitrary compute and storage technology!
- Self-aware instead of unconscious
- Improving over time (self-learning, hardware upgrades)
- Enhanced monitoring

NGE

Community Strategy via a Forum / Open Board



Selected Small Long-Term Projects

University of Reading

I/O Modeling and Diagnosing Causes

- Predict likely reason/cause-of-effect of I/O by just analyzing runtime
- Estimate best-case time, if optimizations would work as intended
- Create a tool that automatizes the process...

Personalized Learning

- Use machine learning to personalize learning of the C online course
- Identify good lections, prescribe the order for students

Appendix

The Performance Challenge

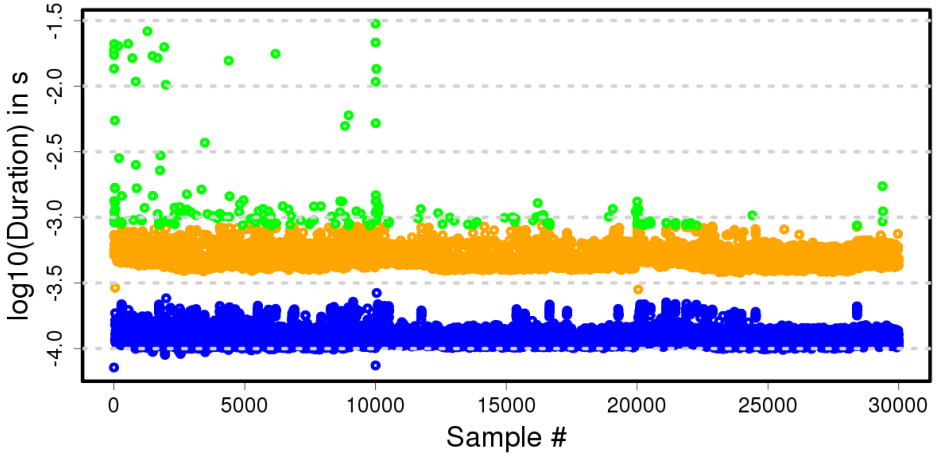
University of Reading

- DKRZ file systems offer about 700 GiB/s throughput
  - However, I/O operations are typically inefficient: Achieving 10% of peak is good
- Influences on I/O performance
  - Application's access pattern and usage of storage interfaces
  - Network congestion
  - Slow storage media (tape, HDD, SSD)
  - Concurrent activity – shared nature of I/O
  - Tunable optimizations deal with characteristics of storage media
  - These factors lead to complex interactions and non-linear behavior

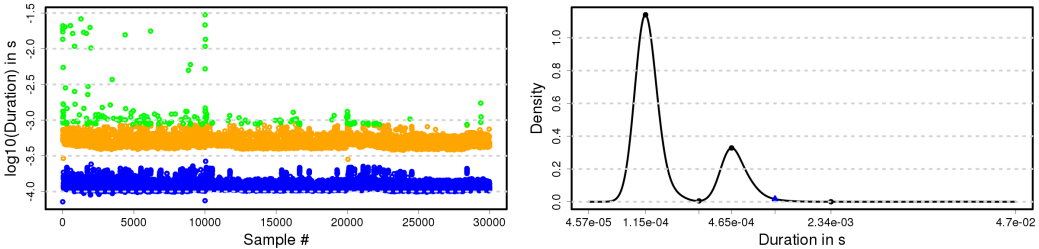
Illustration of Performance Variability

University of Reading

- Rerunning the same operation (access size, ...) leads to performance variation
- Individual measurements – 256 KiB sequential reads (outliers purged)



Comparing Density Plot with the Individual Data Points

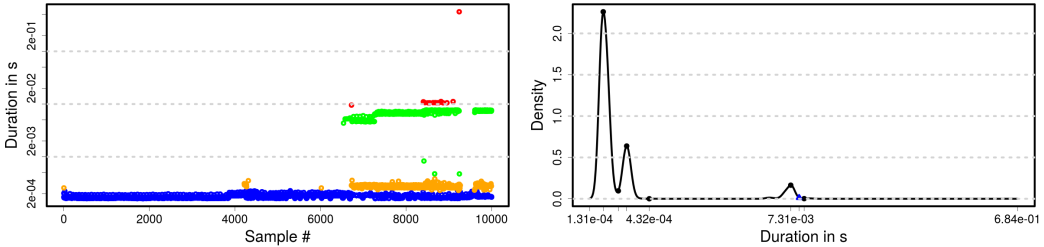


Duration for sequential reads with 256 KiB accesses (off0 mem layout)

Algorithm for determining classes (color schemes)

- Create density plot with Gaussian kernel density estimator
- Find minima and maxima in the plot
- Assign one class for all points between minima and maxima
- Rightmost hill is followed by cutoff (blue) close to zero $\Rightarrow$ outliers (unexpected slow)

Write Operations



Results for one write run with sequential 256 KiB accesses (off0 mem layout).

Known optimizations for write

- Write-behind: cache data first in memory, then write back
- Write back is expected to be much slower

This behavior can be seen in the figure !

Outline

University of Reading

Compression Research: Involvement

University of Reading

- Development of algorithms for lossless compression
  - MAFISC: suite of preconditioners for HDF5, aims to pack data optimally
    Reduced climate/weather data by additional 10-20%, simple filters are sufficient
- Cost-benefit analysis: e.g., for long-term storage MAFISC pays of
- Analysis of compression characteristics for earth-science related data sets
  - Lossless LZMA yields best ratio but is very slow, LZ4fast outperforms BLOSC
  - Lossy: GRIB+JPEG2000 vs. MAFSISC and proprietary software
- Development of the Scientific Compression Library (SCIL)
  - Separates concern of data accuracy and choice of algorithms
  - Users specify necessary accuracy and performance parameters
  - Metacompression library makes the choice of algorithms
  - Supports also new algorithms
  - Ongoing: standardization of useful compression quantities
- A **method for system-wide determination** of data characteristics
  - Method has been integrated into a script suite to scan data centers

Ongoing Activity: Earth-Science Data Middleware

University of Reading

- Part of the ESiWACE Center of Excellence in H2020
  - Centre of Excellence in Simulation of Weather and Climate in Europe
- ESiWACE2 follow up has been funded!

ESDM provides a transitional approach towards a vision for I/O addressing

- Scalable data management practice
- The inhomogeneous storage stack
- Suboptimal performance & performance portability
- Data conversion/merging

Earth-System Data Middleware

University of Reading

Design Goals of the Earth-System Data Middleware

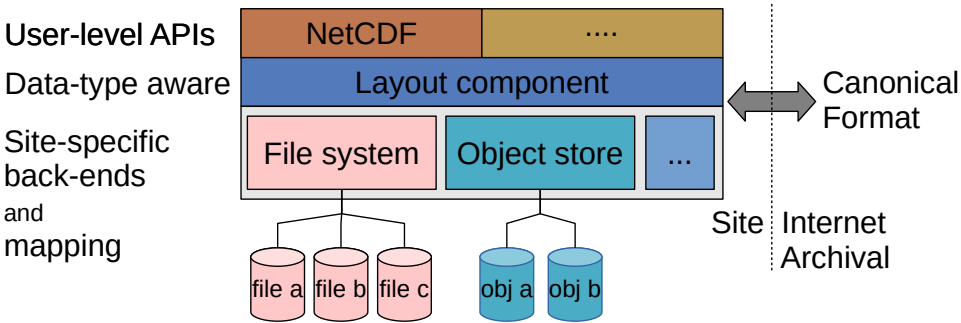
1. Relaxed access semantics, tailored to scientific data generation
   - Avoid false sharing (of data blocks) in the write-path
   - Understand application data structures and scientific metadata
   - Reduce penalties of **shared** file access
2. Site-specific (optimized) data layout schemes
   - Based on site-configuration and performance model
   - Site-admin/project group defines mapping
   - Flexible mapping of data to multiple storage backends
   - Exploiting backends in the storage landscape
3. Ease of use and deployment particularly configuration
4. Enable a configurable namespace based on scientific metadata

Architecture

University of Reading

Key Concepts

- Middleware utilizes layout component to make placement decisions
- Applications work through existing API (currently: NetCDF library)
- Data is then written/read efficiently; potential for optimization inside library



Challenges

University of Reading

- Achieving high performance
- Understanding observed behavior (and performance)
- Tuning system settings and configurations
- Enabling performance portability
- Managing files and (data-intense) workflows
- Utilizing heterogeneous storage landscapes

These are opportunities for tools and method development!

- Diagnosing causes, predicting performance, prescribing settings
- Smarter ways of data handling

Performance Analysis

University of Reading

Problem

Assessing observed time for I/O is difficult.

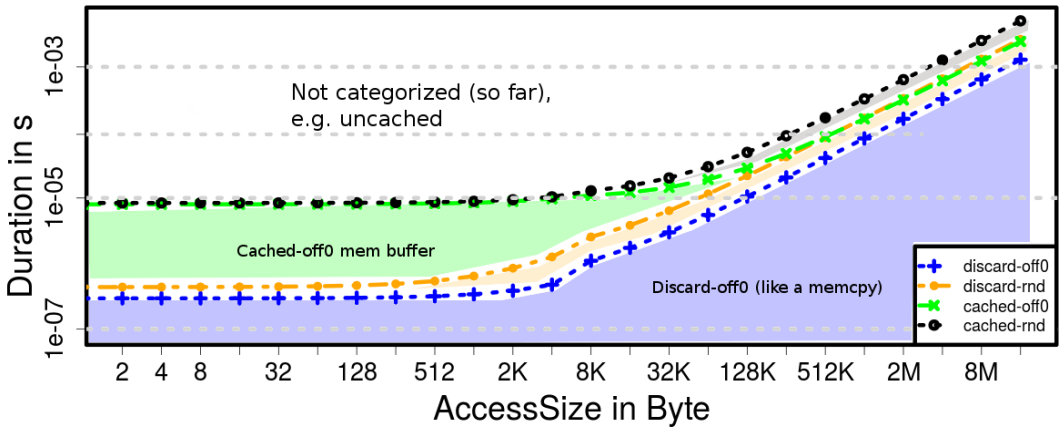
What best-case performance can we expect?

Support for analysis – my involvement

- Models and simulation
  - Trivial models: using throughput + latency
  - PIOSimHD: MPI application + storage system simulator
- Tools to capture and analyze system statistics and I/O activities
  - HDTrace – tracing tool for parallel I/O (+ PVFS2)
  - SIOX – tool to capture I/O on various levels
  - Grafana – Online monitoring for DKRZ (support)
- Benchmarks – on various levels, e.g., Metadata (md-workbench, IOR)
- **Statistic model** to determine likely cause based on time

Resulting Performance Models for Read

University of Reading



Read models predicting caching and memory location.

Using the Model to Identify Anomalies

University of Reading

Using the model, the figure for reverse access shows slow-down (by read-ahead)