ECMWF’s Exascale IO challenges

From inside the HPC to a whole Data archive migration

T. Quintino, S. Smart, O. Iffrig, J. Hawkes, J. Hanley, N. Manubens, E. Danovaro, A. Bonnani, D. Sarmani, B. Raoult, P. Bauer

ECMWF
tiago.quintino@ecmwf.int

SIGIO/UK
Workshop on Storage Challenges in the UK
23rd April 2020
ECMWF’s Forecasting Systems

- Established in 1975.
- Intergovernmental Organisation
  - 22 Member States | 12 Cooperation States
  - 350+ staff
- 24/7 operational service
  - Operational NWP centre
  - Supporting NWS (coupled models) and businesses
- Research institution
  - Closely connected with researchers worldwide
- Operates two Copernicus Services
  - Climate Change Service (C3S)
  - Atmosphere Monitoring Service (CAMS)
- Supports Copernicus Emergency Management Service (CEMS)
ECMWF’s Production Workflow

Global Observations

Acquisition → IFS Model → Product Generation → Product Dissemination

- MARS
- Perpetual Archive
- Obs
- Fields
- Products

Member States & Customers

ECMWF

EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS
ECMWF’s Production Workflow

IFS Model

Product Generation

Dissemination

Raw Output

Parallel Filesystem Storage (Lustre)

Fields

70% Read

Products

Time critical path = 1 hour window

Member States & Customers

Perpetual Archive

MARS

ECMWF
EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS
## Effects of Product Generation using Parallel Filesystem

<table>
<thead>
<tr>
<th></th>
<th>IFS Model (No I/O)</th>
<th>IFS Model + I/O</th>
<th>IFS Model + I/O + PGen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>2440</td>
<td>2776</td>
<td>2926</td>
</tr>
<tr>
<td>Run time [s]</td>
<td>5765</td>
<td>6749</td>
<td>7260</td>
</tr>
<tr>
<td>Relative</td>
<td>-</td>
<td>+ 17%</td>
<td>+ 26%</td>
</tr>
</tbody>
</table>

- **Runtimes affected by the existence of another parallel job in the system:**
  - Product Generation reading the data the model is writing
  - “Coupling” via the file system!

- **9Km 50 member ensemble**
- **Broadwell nodes 2x18 cores**
- **Cray XC40 Aries interconnect**
- **Lustre FS IOR 90GiB/s**
Storage View of Workflow

Observations

Acquisition

Acquire

IFS Model

Parallel FS

Produce

Product Dissemination

Disseminate

Member States & Customers

Archive

MARS

Perpetual Archive

Data is Central!

Product Generation

Modify
NEW HPC Facility + New HPC system

Historical Growth of Generated Products

Model Output Projected Growth
## History and Future of Resolution Upgrades

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Grid size</th>
<th>Grid Points</th>
<th>Field Size (in memory)</th>
<th>Vertical Levels</th>
<th>YEAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>T319</td>
<td>62.5 km</td>
<td>204 k</td>
<td>1.6 MB</td>
<td>L31</td>
<td>1998</td>
</tr>
<tr>
<td>T511</td>
<td>39 km</td>
<td>524 k</td>
<td>4 MB</td>
<td>L60</td>
<td>2000</td>
</tr>
<tr>
<td>T799</td>
<td>25 km</td>
<td>1.2 M</td>
<td>9.6 MB</td>
<td>L91</td>
<td>2006</td>
</tr>
<tr>
<td>T1279</td>
<td>16 km</td>
<td>2.1 M</td>
<td>16.8 MB</td>
<td>L91</td>
<td>2010</td>
</tr>
<tr>
<td>Tco1279</td>
<td>9 km</td>
<td>6.6 M</td>
<td>50.4 MB</td>
<td>L137</td>
<td>2016</td>
</tr>
<tr>
<td>Tco1999</td>
<td>5 km</td>
<td>16.1 M</td>
<td>122.6 MB</td>
<td>L160</td>
<td>2025</td>
</tr>
<tr>
<td>Tco3999</td>
<td>2.5 km</td>
<td>64 M</td>
<td>490 MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tco7999</td>
<td>1.25 km</td>
<td>256 M</td>
<td>1909 MB</td>
<td>L180</td>
<td>2030</td>
</tr>
</tbody>
</table>
TCo7999 (~1.25km) 256 Megapixel

(12 h forecast, *hydrostatic*, no deep convection parametrization, 120s time-step, 960 Broadwell nodes, ~10s per timestep)
Storage and I/O @ Exascale
How large is a 1.25 km x50 ensemble forecast?

- 50 member ensemble forecast
- *Compressed* GRIB2 data @ 16bit & 24bit
- @ 18km O640 → 21 TiB
- Resolution @ 9km O640 → O1280 x 3.3
- Resolution @ 5km O1280 → O1999 x 3.3
- Upgrade levels 137 → 200 x 1.46
- Resolution @ 2.5km O1999 → O3999 x 3.3
- Resolution @ 1.25km O3999 → O7999 x 3.3

21 TiB x 173.2 = 3638 TiB
NextGenIO Prototype

• Read all @ www.nextgenio.eu
• Development of an HPC node by with Intel Optane DCPMM
• Dual-CPU Intel® Xeon® SP nodes (48 cores)
• OmniPath network
• 192GB DRAM
• 3TiB of NVRAM DIMMs (max 6 TiB)

• Prototype system
  – 34 compute nodes
  – Hosted @ EPCC, Edinburgh

34 x 3 TiB Byte Addressable Storage
FDB (version 5)

- Domain specific (NWP) Distributed object store
- Transactional, No synchronization
- Key-value store
  - Keys are scientific meta-data (MARS Metadata)
  - Values are byte streams (GRIB)

- Support for multiple back-ends:
  - POSIX file-system (currently on Lustre)
  - NVRAM using PMDK library

- Supports wild card searches, ranges, data conversion, etc...

param=temperature/humidity, levels=all, steps=0/240/by/3, date=01011999/to/31122015,
FDB 5 Semantics

1. ACID – *Transactional*
2. Write blocks until data handed over – *Asynchronous*
3. `flush()` blocks until data is visible – *Consistent*
4. Write-once, don’t overwrite - *Immutable*
5. Data can be masked – *Versioned*

- All I/O operations are asynchronous, so computation can continue
- Distributed to all servers using a *Rendezvous Hash*, so no synchronisation needed
Front-ends and API

- Determines where the data is stored …
  - Run-time configurable
  - Implement data collocation policies
  - Manage data pools
  - Implements a simple interface:

```
archive(Metadata key, void* data, size_t length);
retrieve(Metadata key, void* data, size_t& length);
flush();
```

Metadata:
CLASS = OD, TYPE = FC, LEVTYPE = PL, EXPVER = 0001, STREAM = OPER, PARAM = 130, TIME = 1200, LEVELIST = 500, DATE = 20190614, STEP = 12
FDB5 Data Routing

- Meta-data controlled routing
- Fully asynchronous I/O
- Remote access TCP/IP

IFS Model

Select

Distribution

Remote

Storage backend
Multiple hosts

class=od

class=rd
Asynchronous Archiving Data

**Client**
- `archive()`
- `flush()`
- Push data onto queue
- Pop data off queue.

**Server**
- `archive()`
- `flush()`
- Starts threads to:
  - Receive data
  - Archive data

1. **Archive**
2. **Acknowledge**
3. **Data blobs**
4. **Flush**
5. **Acknowledge**

Push data onto queue

Pop data off queue.

Forward to API

Wait until all archives complete
% fdb stats class=od,date=20190612,expver=0001
Summary:
========
Number of databases             : 58
Fields                          : 83,747,723
Size of fields                  : 104,493,002,498,506 (95.0358 Tbytes)
Duplicated fields               : 1,316,502
Size of duplicates              : 2,668,035,857,106 (2.42656 Tbytes)
Reachable fields                : 82,431,221
Reachable size                  : 101,824,966,641,400 (92.6093 Tbytes)
Databases                       : 58
TOC records                     : 89,329
Size of TOC files               : 191,427,584 (182.56 Mbytes)
Size of schemas files           : 949,228 (926.98 Kbytes)
TOC records                     : 89,329
Owned data files                : 89,271
Size of owned data files        : 104,506,303,059,882 (95.0479 Tbytes)
Index files                     : 89,271
Size of index files             : 13,677,232,128 (12.7379 Gbytes)
Size of TOC files               : 191,427,584 (182.56 Mbytes)
Total owned size                : 104,520,172,668,822 (95.0605 Tbytes)
Total size                      : 104,520,172,668,822 (95.0605 Tbytes)
FDB 5 Parallel Write Performance to DCPMMs

- Application data measured
- Consistency semantics
- Includes wait for $\text{flush}()$
FDB 5 Parallel Read Performance to DCPMMs

- Read slower than write
- Includes the *data lookup* in the indexes

Data read rate (GiB/s) vs. Number of reader processes for different numbers of server nodes (2, 4, 8, 16, 24). The graph shows a trend where the data read rate increases with the number of reader processes, with a peak at 60 GiB/s for 128 reader processes.
Running the forecast model

<table>
<thead>
<tr>
<th></th>
<th>Model + I/O</th>
<th>Model + I/O + PGen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run time (Lustre) [s]</td>
<td>1793</td>
<td>1928</td>
</tr>
<tr>
<td>Run time (Distributed) [s]</td>
<td>1610</td>
<td>1599</td>
</tr>
</tbody>
</table>

Runtimes no longer affected by the Product Generation!!!

NextGenIO prototype. 32 nodes
Intel OmniPath2 interconnect
6 ensemble members
Preliminary Results

ECMWF Operational Filesystem

- Sonexion snx11061
- OST Nodes: 288
- 20TiB per node (10 disks)
- 4PiB capacity
- Measured 165GiB/s (IOR)

- Sustained IFS runs: R 22.4 GiB/s + W 22.0 GiB/s = 44.4 GiB/s application data

NEXTGenIO + Distributed FDB

- Nodes: 34
- 3TiB per node (12 DIMMs)
- 108 TiB capacity

- Not yet optimised!
- Measured sustained 72 GiB/s W application data (16 nodes)
Can we handle the 1.25 km ensemble forecast?

- 50 member ensemble forecast
- *Compressed* GRIB2 data @ 16bit & 24bit
  - @ 1.25km 7999, 3638 TiB
  - Required to read 70%, x 1.70
  - @ 1.25km 7999, 6185 TiB
  - Time to solution 1 hour 6185 TiB / 3600 = 1759 GiB/s
  - NextGenIO performance (16 nodes), 132 GiB/s
  - Required Nb Prototypes 1759 / 132 x 16 = 213 nodes

NextGenIO x 6.7 (by 2035)
ECMWF Novel Data Flows

Data Analytics / Machine Learning

Producer

FDB Data Hypercubes

Consumer

- Bring users to the data
- Use data while it is hot
- Access using scientifically meaningful metadata

Archive

PFS

Cloud

HDD

Tape

MARS
Providing ECMWF Data to a Cloud

Requirements:
1. Bring users to the data and avoid moving the data out of the data centre
2. Provide users with computing resources collocated with data
3. Data-centric approach "move the compute, not the data"

How to enable this:
1. Mechanism to pull/push data from ECMWF
2. Mechanism to run custom post-processing
3. Mechanism to explore & discover data

New development: Polytope
Watch this space ;-)
Messages To Take Home

*Ensemble data sets are growing quadratically to cubically in size, How can we best serve this high-resolution data?*

*New technologies in the horizon*  
**NVRAM and other Storage Class Memories**

*ECMWF is adapting its workflow to take advantage of these upcoming technologies*  
*Developed a distributed object store for Weather and Climate*  
*Working to serve these datasets out of the HPC to Data Analytics Platforms*
How about that move of Data Centre?
ECMWF’s Production Workflow

Global Observations

200 GiB/day

Acquisition

250 TiB/day

MARS

IFS Model

100 TiB/day

Product Generation

70 TiB/day

Product Dissemination

30 TiB/day

Member States & Customers

Productions

Obs

Fields
Moving a Data Centre

How to move a 24x7 data center?

- **Run weather forecast 4x per day**
- **Still produce ~ 100TiB/day**
- **Obtain a new HPC** and install in place

- **Main issue is Data Handling System (DHS)**
  - **350 PiB growing @ 1PiB / 4 days**

**Transfer?**

350 PiB @ 100Gbips network = 339 days

350 PiB @ 300TiB/day tape access = 1194 days
Code Digression

- *How to do a multi-threaded transactional swap …*

\[
S = 350 \\
B = 0
\]

\[
\text{Lock}(S), \text{Lock}(B)
\]

\[
\begin{align*}
\text{TMP} &= S \\
S &= B \\
B &= \text{TMP}
\end{align*}
\]

\[
\text{Unlock}(B), \text{Unlock}(S)
\]

\[
\text{clean}(\text{TMP})
\]
**DHS Service Transition plan**

**SHINFIELD**

- **Original Archive off.**
- **Transfer Archive DBs to Bologna**

**ARCHIVE DARK PERIOD**

- **Flush disks to tape**
- **Transfer HPSS index DBs**

**BOLOGNA**

- **Dry-runs**
- **Temporary copy (write) in production (3 months)**

- **Temporary copy (read)**
- **Temporary master (write/read)**

- **MARS+ECFS+HPSS Bologna ON**
  - empty disks
  - blank tapes

- **Dismantle tape libraries and disks for physical transfer to Bologna**

- **DHS fully loaded**

**D**

- **D+4**
- **D+6**

**DHS**

**EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS**
THANK YOU!

QUESTIONS?

This work has been supported by NextGenIO project and partly funded by European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement 671951

This work has been supported by LEXIS project and partly funded by European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement 824115