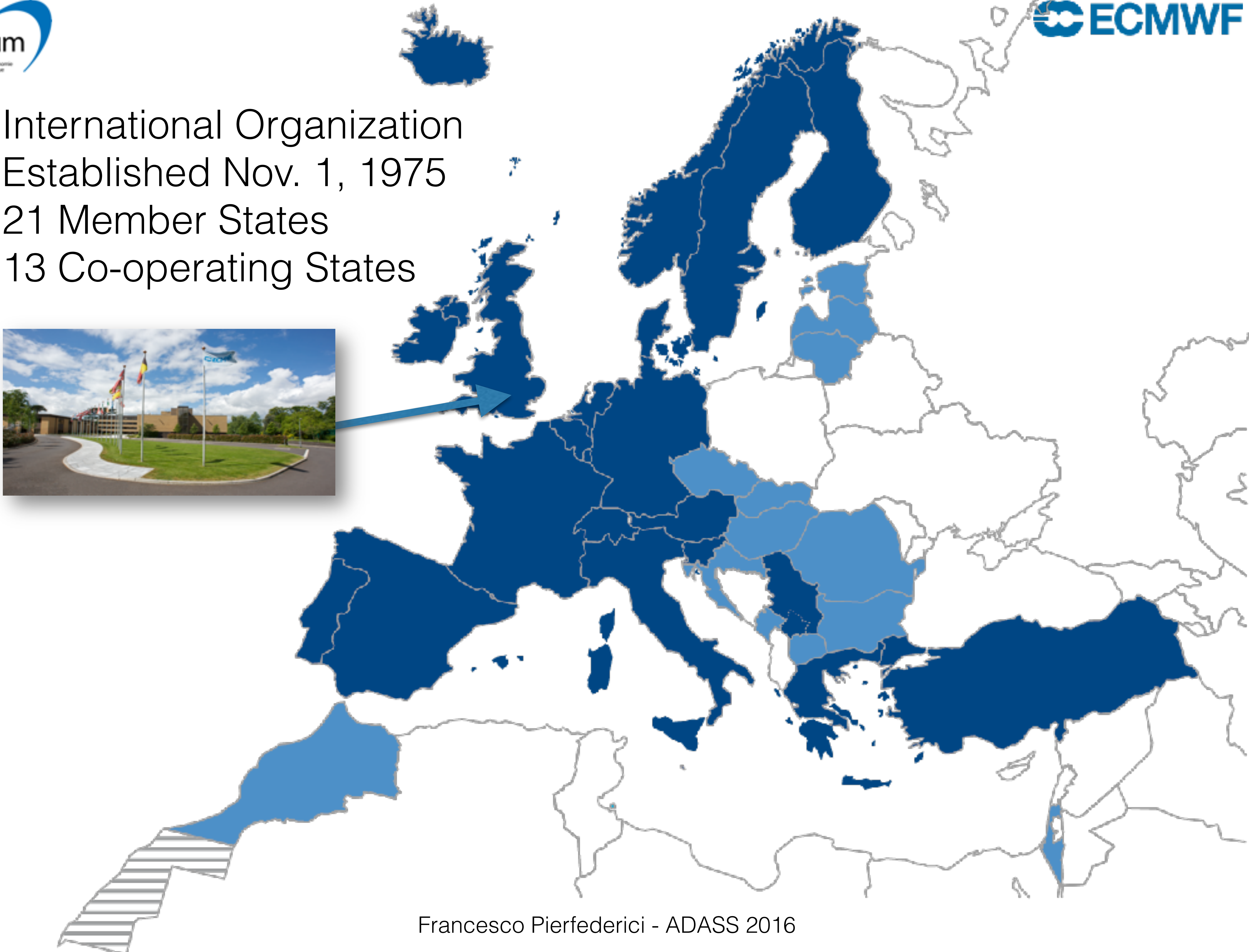
Massive Scientific Workloads

Lessons Learned From Petaflop-Scale Weather Simulations

Francesco Pierfederici

International Organization
Established Nov. 1, 1975
21 Member States
13 Co-operating States



Operational Forecast

- High resolution deterministic forecast: twice per day - 9 km 137-level, to 10 days ahead

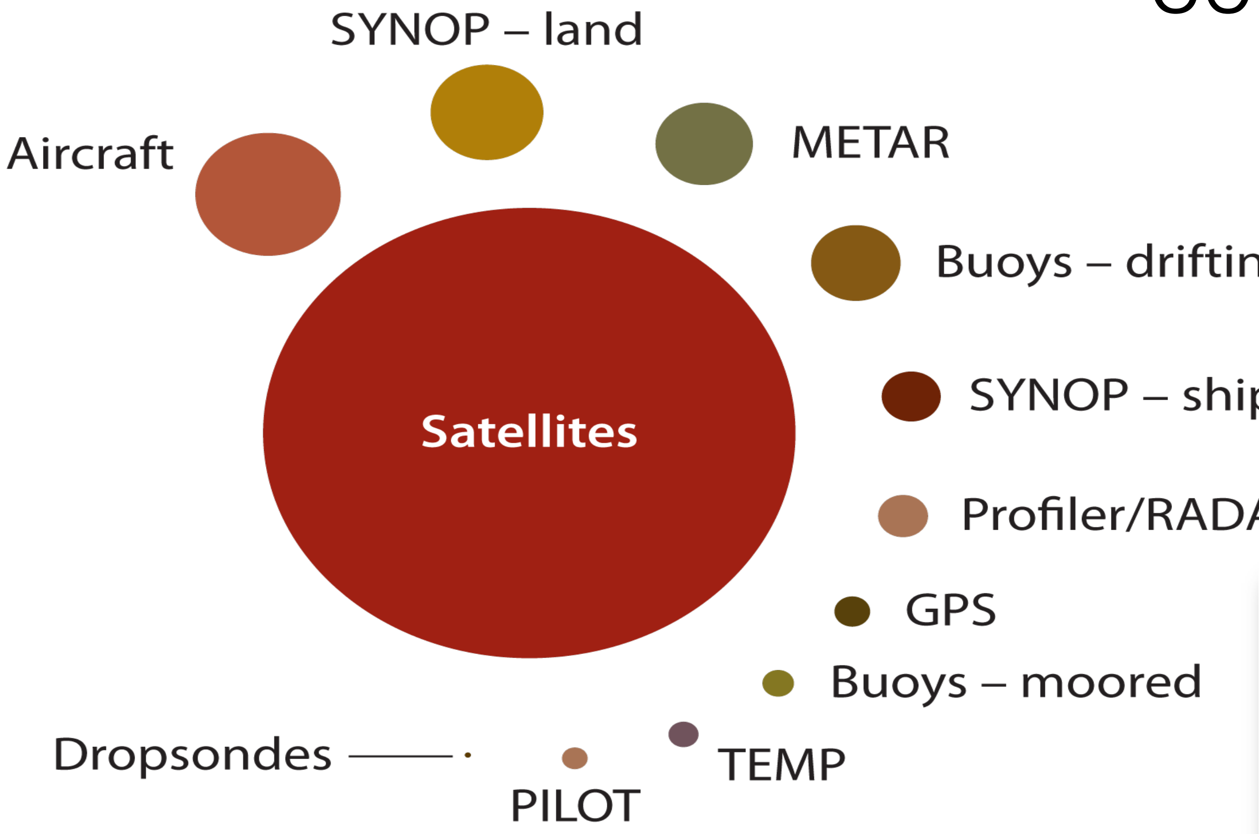
- Ensemble forecast: twice daily - 51 members, 20/30 km 91-level, to 15 days ahead

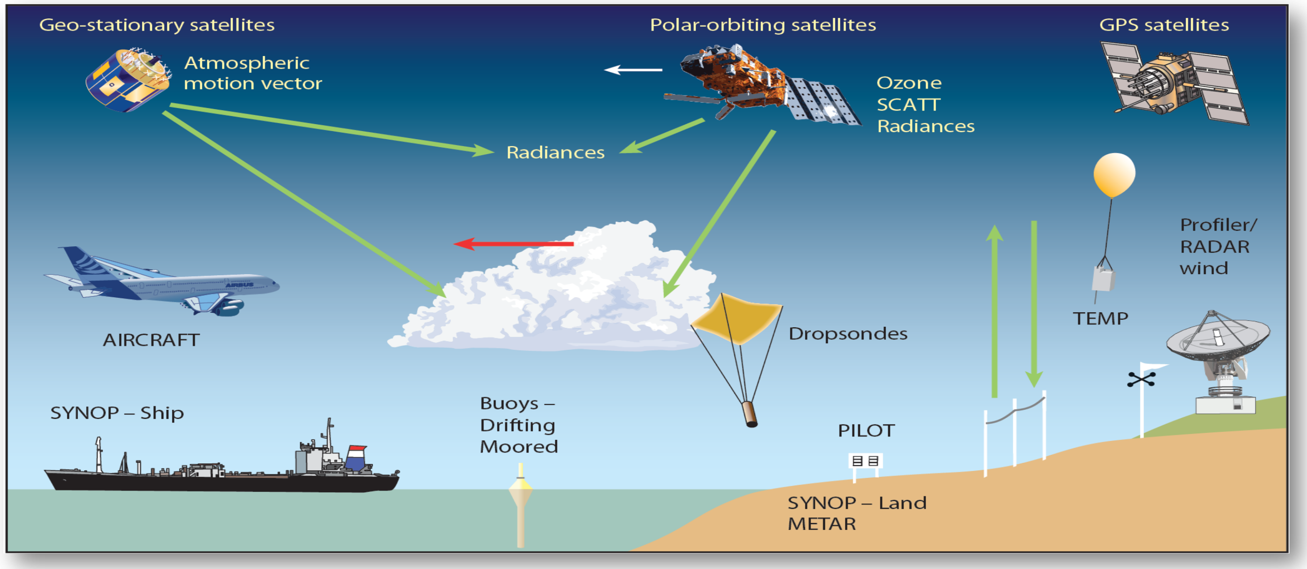
- Monthly forecast: twice a week - 51 members, 20/30 km 91 levels, to 1 month ahead (46 days ahead)

- Seasonal forecast: once a month - 51 members, ~80 km, 91 levels, to 7 months ahead

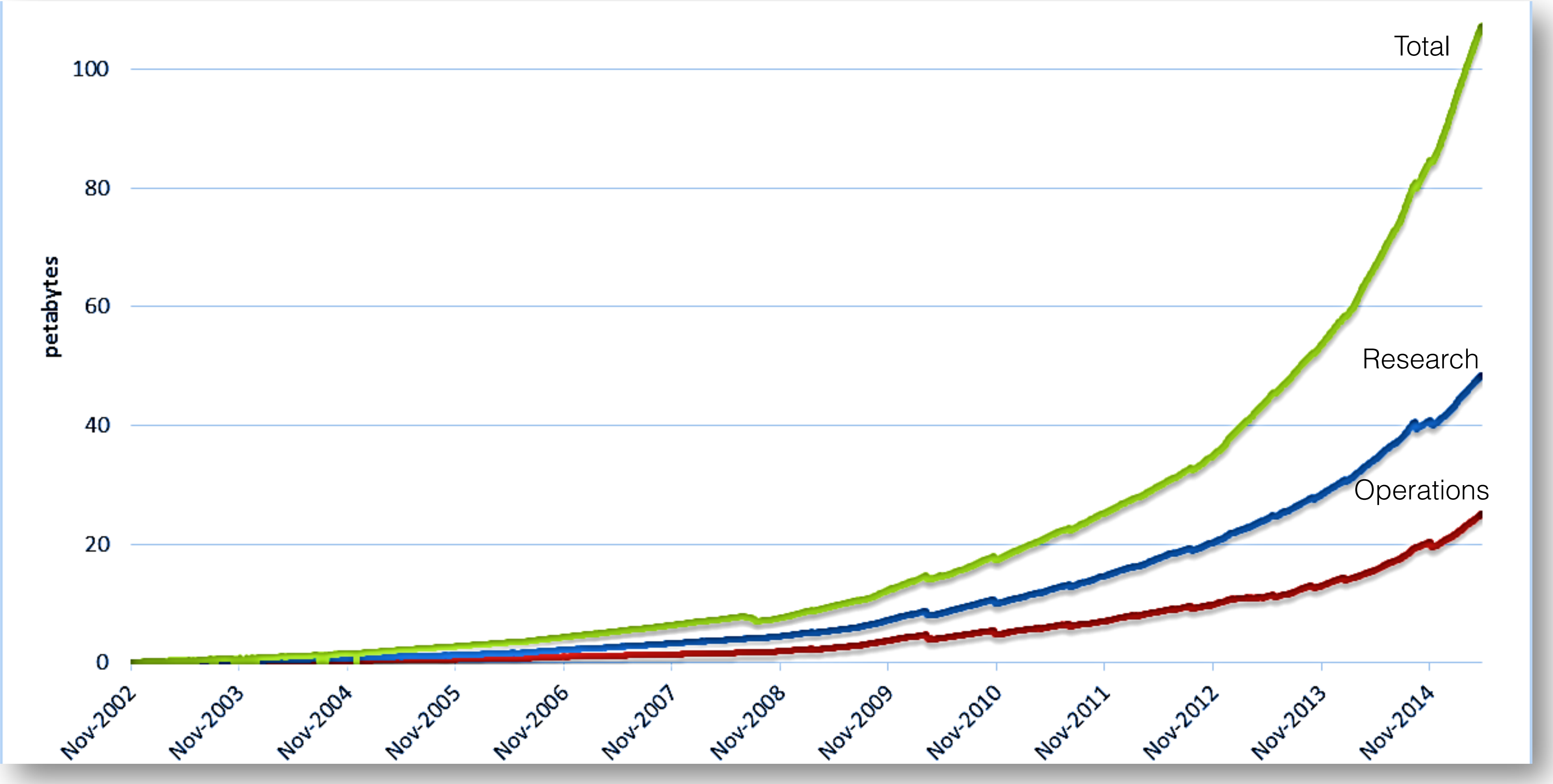
Forecast Data

~60×10⁶ observations / 12 hours





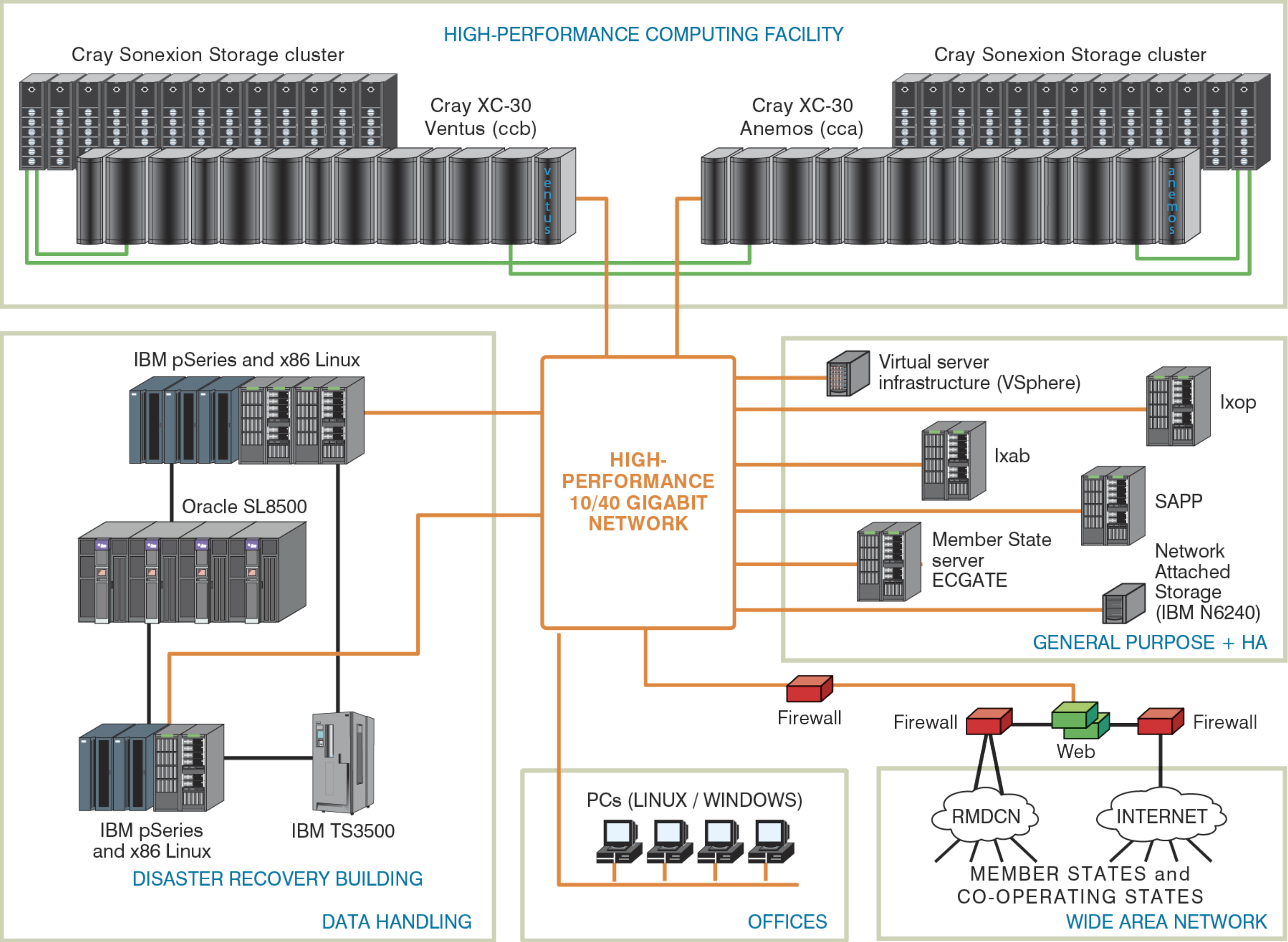
The Archive



Yearly growth rates between 37% - 58% depending on HPC availability (~ 1PB/week at the moment)

Our HPC

Sustained ECMWF Code	333 Teraflops
Peak	8500 Teraflops
Clusters	2
Compute Nodes	7,220
Compute Cores	259,920
Pre/Post Proc Nodes	128
Cores/node	18 x 2
Memory	920 TB
Lustre FS	> 20PB
Interconnect	Cray Aries (16GB/s/dir)
Scheduler	PBS Pro



HIGH-PERFORMANCE COMPUTING FACILITY

Cray Sonexion Storage cluster

Cray XC-30
Ventus (ccb)

Cray Sonexion Storage cluster

Cray XC-30
Anemos (cca)

IBM pSeries and x86 Linux

Oracle SL8500

IBM pSeries
and x86 Linux

IBM TS3500

DISASTER RECOVERY BUILDING

DATA HANDLING

HIGH-PERFORMANCE 10/40 GIGABIT NETWORK

Virtual server
infrastructure (VSphere)

lxop

lxab

SAPP

Member State
server
ECGATE

Network
Attached
Storage
(IBM N6240)

GENERAL PURPOSE + HA

Firewall

Firewall

Web

Firewall

PCs (LINUX / WINDOWS)

OFFICES

RMDCN

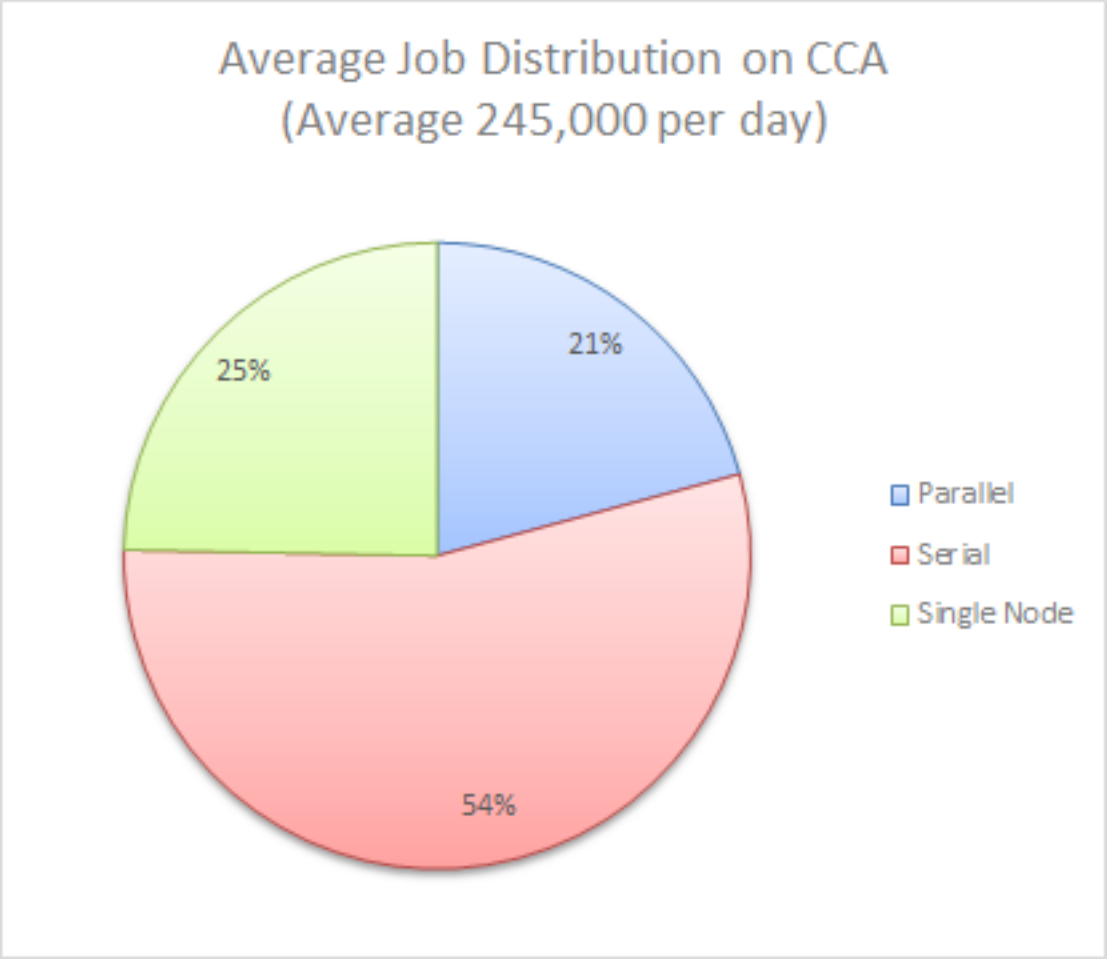
INTERNET

MEMBER STATES and
CO-OPERATING STATES

WIDE AREA NETWORK

Experiments

- Forecast/reanalysis/climate simulations
- Each Experiment has thousands of steps/tasks
- Each task can be (i.e. often is) an MPI job
- Tasks can be composed into higher-level tasks (families)



Challenges

- Application performance bottlenecks
- Workflow performance bottlenecks (e.g. network contention, lustre performance)
- Power usage and availability

Challenges

- Current tools (e.g Alinea MAP and Darshan) only work at the compiled code level
- No off-the-shelf tools for workflow-level analysis

Insights

- Workflow -> Application
- HPC Cluster -> Single machine
- Power defines the envelope

Ideas

- Predict resource utilisation (including network)
- Interleave computation & IO
- Oversubscribe nodes
- Find the sweet spot in the power/nodes/cores/time space

Measure

- Full workflow profiler (non intrusive, non sampling)
- High-performance (5K-10K hits/second/workflow)
- Generates a model of
    - computation
    - IO
    - communication
- Feeds back to the resource manager

Thank you!