National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

The Met Office's Climate model at AWS

Simon Wilson NCAS-CMS

Thanks to David Wallom and Bryan Lawrence.

Funding from AWS in Research.

Motivation

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

- Modern climate and weather models are massively parallel and computationally expensive, so tend to be run on supercomputers and clusters.
- Virtual clusters in clouds have been proposed as an alternative HPC platform for running weather and climate models.
- Project investigated running the Met Office's climate model, HadGEM3-A at AWS (Amazon Web Services).
- Direct comparison with systems currently available to the scientific community using standard model configurations.
- **Very much a work in progress.**
- Hope soon to run on Azure HPC.

Requirements

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

- More than just running a climate model on the cloud. If the cloud is to be a viable alternative to “traditional” systems, various aspects need to be considered.
- Speed
- Local I/O
- Data storage, ingress and egress
- Ease of use
- Cost

This talk is mainly about speed, scaling and compute costs. Other aspects will be mentioned.

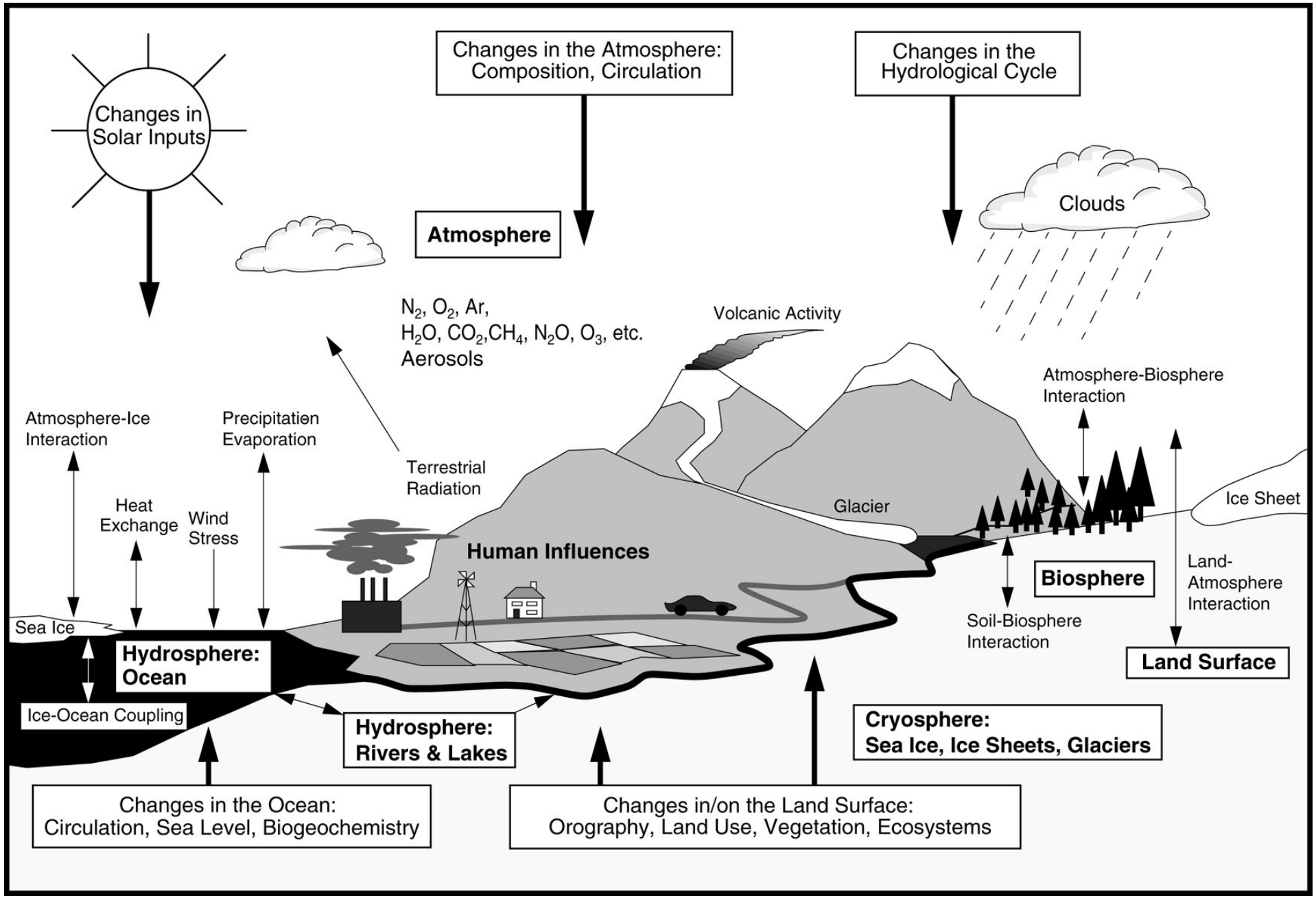
What is a Climate Model

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

- Mathematical model of the earth system which simulates weather over long time scales.
- Similar formulation as models used for forecasting but with the addition of longer timescale processes.
- Very complex, with many different types of physical processes simulated.
- Generally run for 10s or 100s of model years.
- Computationally very expensive.
- TBs of output data.
- Somewhere between .5 and 1 million lines of FORTRAN using MPI for communications between cores.
- Running on 10,000s of cores possible.

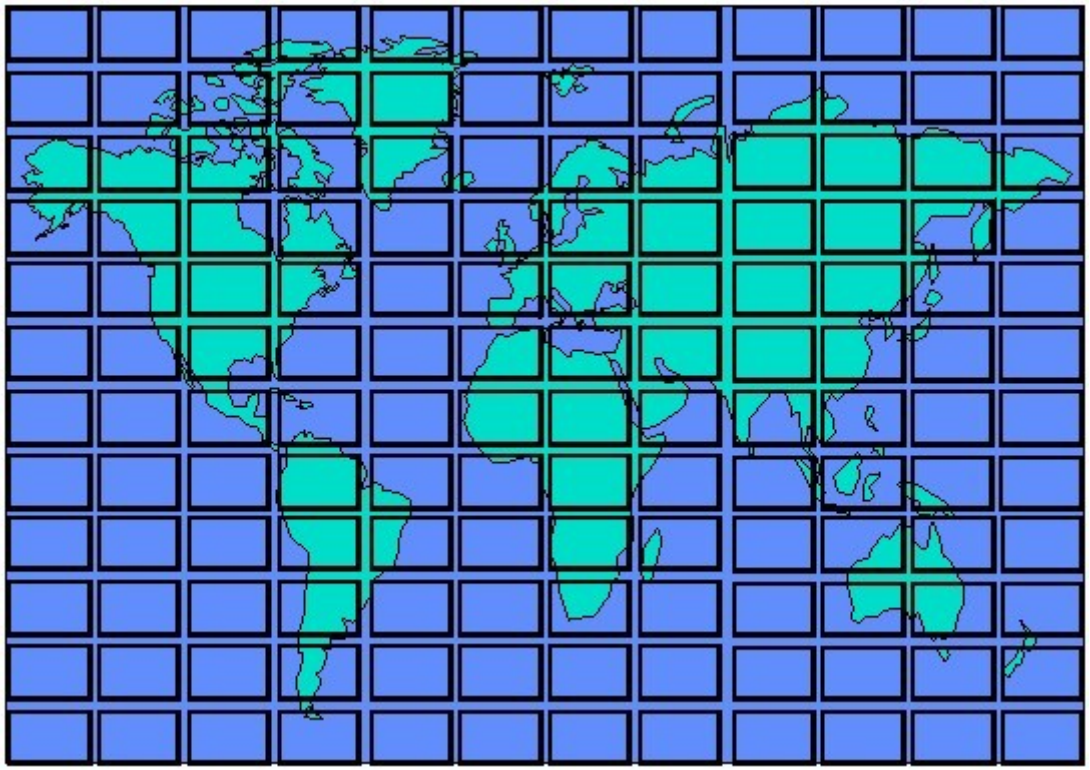
Processes within a Climate Model

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



Model decomposition

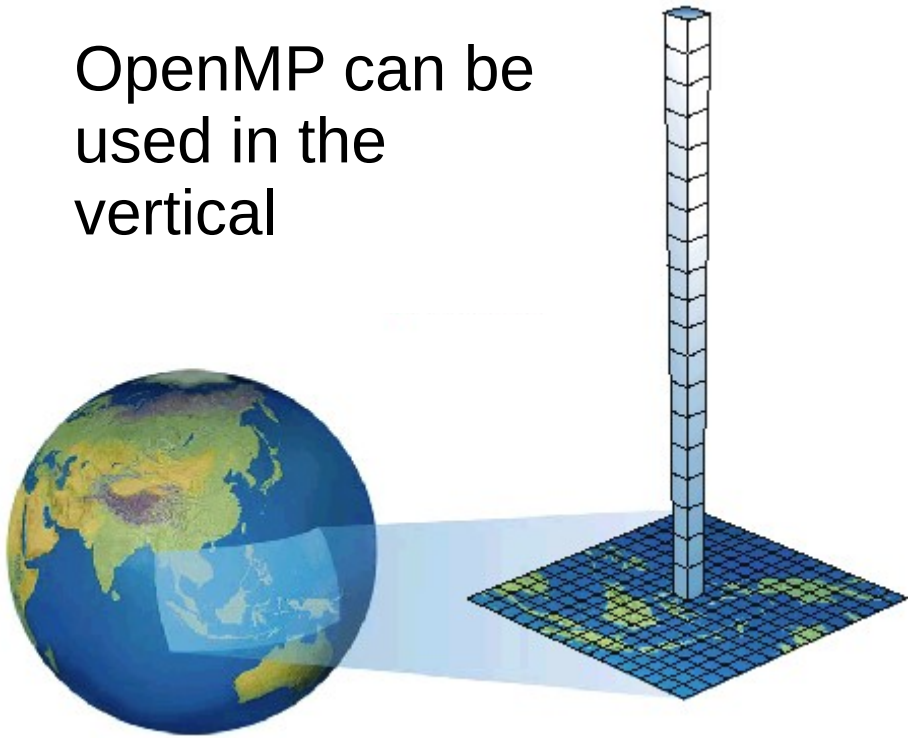
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



The model is massively parallel. The globe is divided up into grid boxes which are distributed over the cores.

Relies on MPI for message passing between cores. As core number increases as does the amount of communication. Mainly halo exchange, all to one gathers/scatters and reductions.

OpenMP can be used in the vertical



Systems configuration

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Archer Cray XC30 2 2.7GHz, 12-core E5-2697 v2 (Ivy Bridge) Intel per node

64GB memory per node
Aries interconnect. Fast, low latency, high bandwidth.
Intel Fortran v15
Cray MPICH2

Lotus 2 2Ghz 8 core Intel Xeon E5-2650 v2 “Ivy Bridge” per node

128GB memory per node
10GB interconnect
Intel Fortran v15
IBM platform MPI (MPI2)

AWS 2 2.9GHz(3.5 GHz turbo) 10 core E5-2666 v3 Haswell per node

60GB memory per node
10GB interconnect
c4.8xlarge Amazon Linux VM instance
Intel Fortran v15 standard
MPICH3 (compiled at Reading)

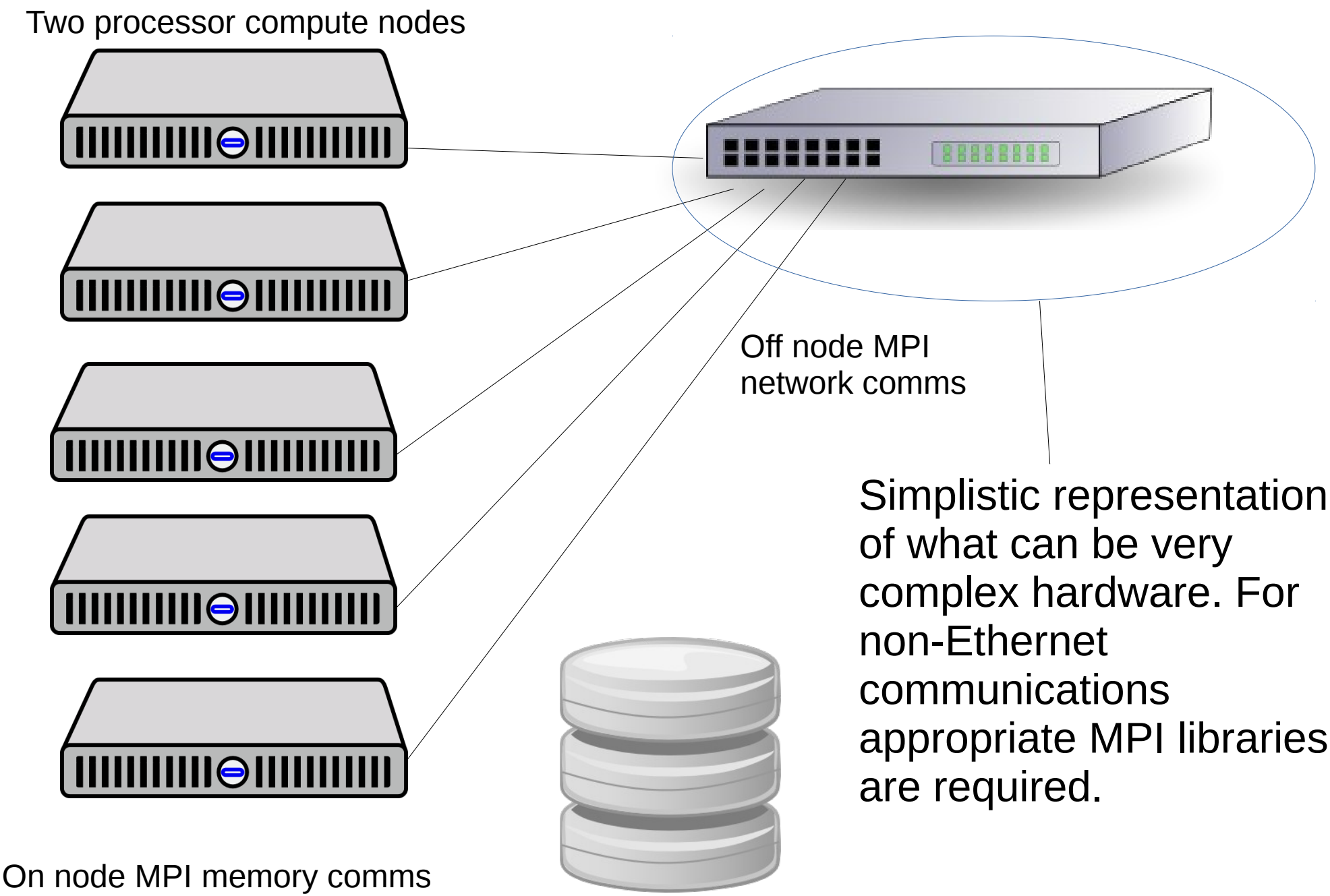
AWS cluster

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

- C4.8large instances 18 cores (36 with hyperthreading). SSDs for storage.
- Custom Amazon Linux Amazon Machine Image (AMI) VM instances with benchmark executable, control system, data and MPICH3 runtime installed.
- Private 10GB Ethernet network using Virtual Private Cloud (VPC), analogous to a private network on a cluster.
- Maximum size tested: 32 VMs, 576 cores.
- Standard MPICH3 sockets over Ethernet.
- Benchmark executable compiled statically with Intel Fortran using MPICH3 on another system.

Layout schematic as seen by MPI

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



Simplistic representation of what can be very complex hardware. For non-Ethernet communications appropriate MPI libraries are required.

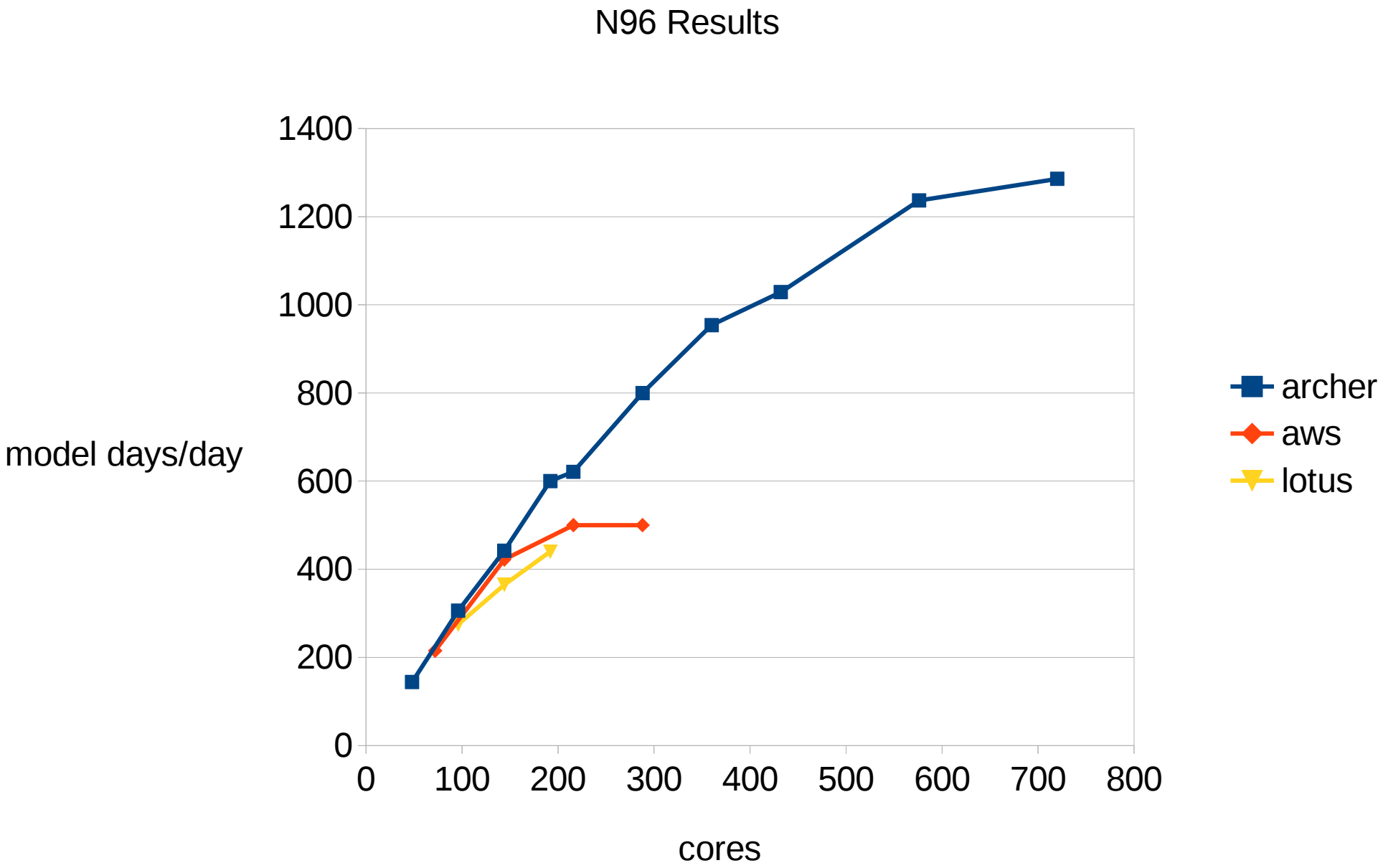
Experimental design

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

- Two configurations chosen N96 and N216.
  - N96 (192x144x85 ($2.4x10^6$) grid points) low resolution, mainly run as an ensemble.
  - N216 (432x342x85 ($1.3x10^7$)) higher resolution, more accurate representation of weather but more computationally expensive.
- Representative of models which will be run over the coming years for the next IPCC assessment report.
- Full diagnostic set. Approximately 1500 separate model fields, mixture of single and multi-level.

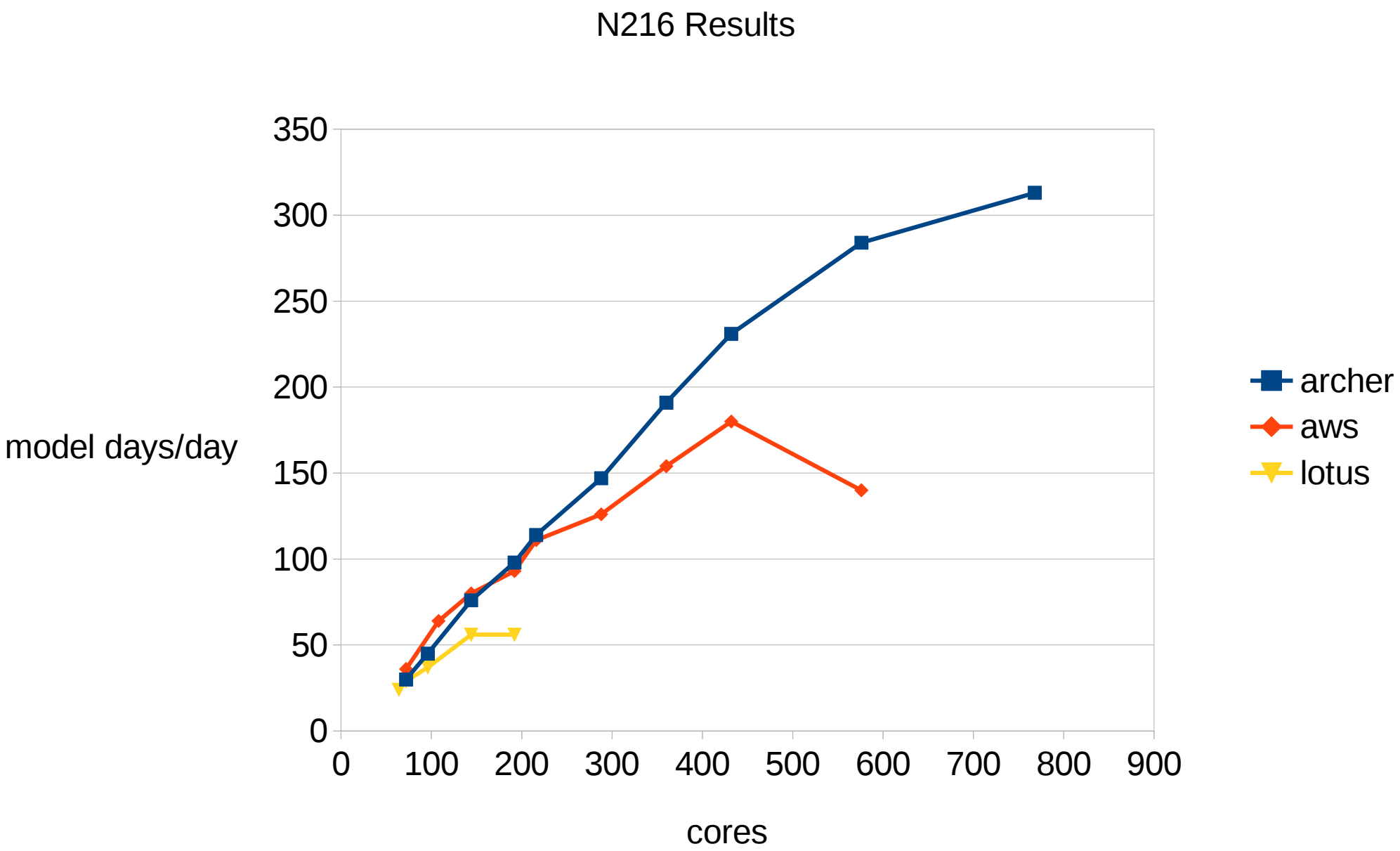
Model run rates

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



Model run rates

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL



Cost analysis of compute

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

- For both Archer and AWS calculated the cost per core per real day.
- Assumed $.40 per hour per node AWS Spot Instance cost.
- On demand cost significantly more.
- Archer calculation of cost per core considered total cost of hardware (including disks), buildings and recurrent costs. Thanks to Bryan Lawrence.
- AWS: £0.43 per core per real day.
- Archer: £0.30 per core per real day.

N96 compute cost (provisional)

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Archer

PEs	Model days/day	Approximate cost per model year (£)
48	144	37
96	306	34
144	442	36
192	600	35
216	621	38
288	800	40
360	954	42
432	1029	46
576	1237	51
720	1286	62

AWS

PEs	Model days/day	Approximate cost per model year (£)
72	215	52
144	422	53
216	500	67
288	500	90

N216 compute cost (provisional)

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Archer

PEs	Model days/day	Approx. cost per model year (£)
72(h)	30	264
96(h)	45	234
144	76	208
192	98	215
216	114	208
288	147	215
360	191	207
432	231	205
576	284	223
768	313	270

(h) high memory node

AWS

PEs	Model days/day	Approx. cost per model year (£)
72	36	310
108	64	261
144	80	279
180	93	300
216	111	301
288	126	354
360	154	362
432	180	372
576	140	637

Data and I/O

N96 144 cores model days/day

	Archer	AWS
No checkpoint/no diagnostics	445	443
Checkpointing/no disgnostics	406	388
No checkpointing/diagnostics	416	422
Checkpointing/diagnostics	370	301

- A N96 model with a full output diagnostic set writes ~150GB diagnostic data per model year, whilst N216 writes 600GB.
- 100 year runs therefore produce 15TB and 60TB respectively
- Can get a constant 150MB/s from Frankfurt AWS to jasmin.
- Input data size ~10GB for N96 and ~40GB for N216 configurations.

Scaling discussion

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

- All three systems have similar performance on processor counts under 250/node count fewer than 15, compute dominates.
- Beyond this scaling is much better on the Cray.
- Due large bandwidth, low latency Ares interconnect (~1μs) on the Cray vs 10GB Ethernet interconnect (~80μs latency) at AWS.
- Archer cost per core per model year less than AWS Spot costs, assuming a Spot price of $.40 per node per hour. At $.30 costs comparable.
- Better scaling on Cray means experiments run faster for same cost.
- Lower resolution models more suited for VMs.

Data discussion

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

- Data egress rates to jasmin are sufficient for output data to be archived directly rather than on S3. Need to calculate egress costs.
- Runtime I/O rates similar to Archer and will have an acceptable impact on run speed.

Future work

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

- Complete analysis.
- Azure HPC. This has a fast Infiniband interconnect which should allow the model to scale to more cores.
- No OpenMP in this study, might improve scaling.
- Integration with standard ROSE/CYLC workflow engine. Already been demonstrated in a Met Office project than it can work with VMs.
- Targeted prebuilt configurations and VMs which will permit the wider research community to use the UM.

Questions