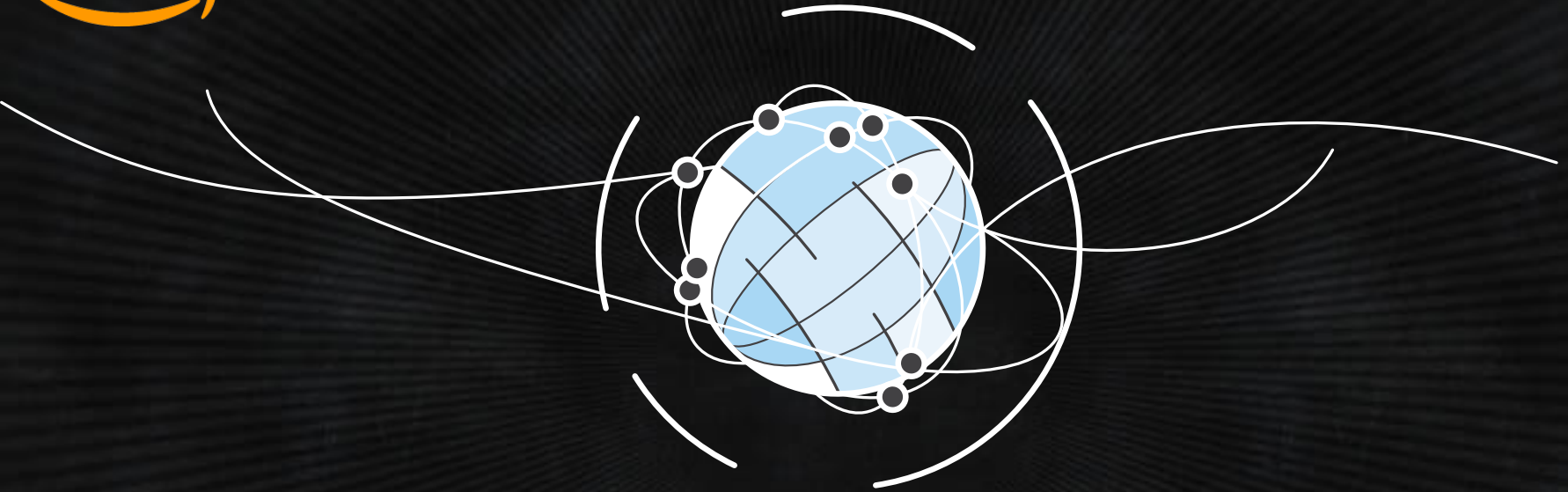


Cloud – The New Frontier of Scientific Research

Vincent Quah
Regional Head – Education, Research, Healthcare and Not For Profit
AWS Asia Pacific and Japan

What research has been successful in the cloud

and why

aws

Key Strengths of AWS for Scientific Discoveries

Time to discovery

- Availability of resources, scalability, right-sizing
- Experiment fast
- Avoid undifferentiated work

Availability of resources: We're off to a cute start …

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0153-5

OPEN

Adaptation and conservation insights from the koala genome

Rebecca N. Johnson[1,2,30,31*], Denis O'Meally[2,3,30], Zhiliang Chen[4,30], Graham J. Etherington[5], Simon Y. W. Ho[2], Will J. Nash[5], Catherine E. Grueber[2,6], Yuanyuan Cheng[2,7], Camilla M. Whittington[8], Siobhan Dennison[1], Emma Peel[2], Wilfried Haerty[5], Rachel J. O'Neill[9], Don Colgan[1], Tonia L. Russell[10], David E. Alquezar-Planas[1], Val Attenbrow[1], Jason G. Bragg[11,12], Parice A. Brandies[2], Amanda Yoon-Yee Chong[5,13], Janine E. Deakin[14], Federica Di Palma[5,15], Zachary Duda[9], Mark D. B. Eldridge[1], Kyle M. Ewart[1], Carolyn J. Hogg[2], Greta J. Frankham[1], Arthur Georges[14], Amber K. Gillett[16], Merran Govendir[8], Alex D. Greenwood[17,18], Takashi Hayakawa[19,20], Kristofer M. Helgen[1,21], Matthew Hobbs[1], Clare E. Holleley[22], Thomas N. Heider[9], Elizabeth A. Jones[8], Andrew King[1], Danielle Madden[3], Jennifer A. Marshall Graves[11,14,23], Katrina M. Morris[24], Linda E. Neaves[1,25], Hardip R. Patel[26], Adam Polkinghorne[3], Marilyn B. Renfree[27], Charles Robin[27], Ryan Salinas[4], Kyriakos Tsangaras[28], Paul D. Waters[4], Shafagh A. Waters[4], Belinda Wright[1,2], Marc R. Wilkins[4,10,30], Peter Timms[29,30] and Katherine Belov[2,30,31]

The koala, the only extant species of the marsupial family Phascolarctidae, is classified as 'vulnerable' due to habitat loss and widespread disease. We sequenced the koala genome, producing a complete and contiguous marsupial reference genome, including centromeres. We reveal that the koala's ability to detoxify eucalypt foliage may be due to expansions within a cytochrome P450 gene family, and its ability to smell, taste and moderate ingestion of plant secondary metabolites may be due to expansions in the vomeronasal and taste receptors. We characterized novel lactation proteins that protect young in the pouch and annotated immune genes important for response to chlamydial disease. Historical demography showed a substantial population crash coincident with the decline of Australian megafauna, while contemporary populations had biogeographic boundaries and increased inbreeding in populations affected by historic translocations. We identified genetically diverse populations that require habitat corridors and instituting of translocation programs to aid the koala's survival in the wild.

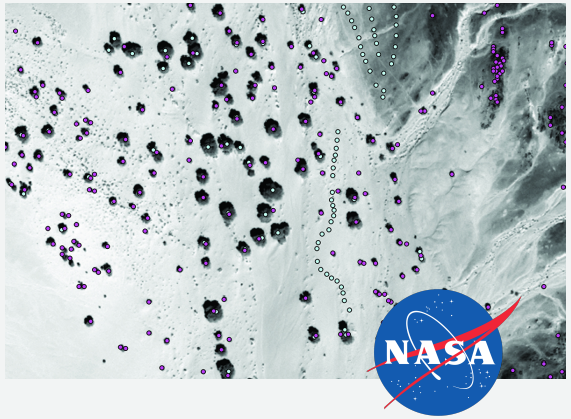
length of the reads at the 60% percentile was calculated as 10,889 bp. The FALCON assembly was run on Amazon Web Service Tokyo region using r3.8xlarge spot instances as compute node, with the number of instances varying from 12 to 20 depending on availability.

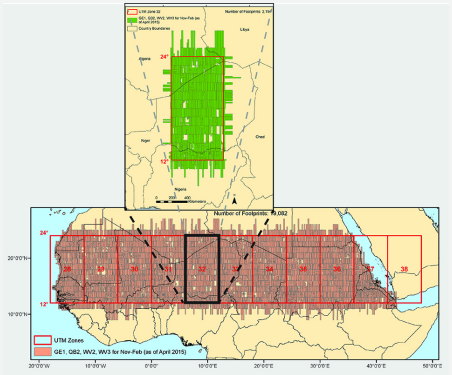
https://aws.amazon.com/blogs/aws/saving-koalas-using-genomics-re

aws

Availability of resources: NASA – Climate Research

- Mosaicking 2,500+ QuickBird satellite images into 100-km x 100-km tiles, which are then broken into 25-km x 25-km sub-tiles for processing.
- Orthorectifying and mosaicking all satellite data in ADAPT
- Identifying trees and shrubs using adaptive vegetation classifier algorithms. Estimating biomass. Incorporating algorithms to calculate tree and shrub height for biomass estimates.





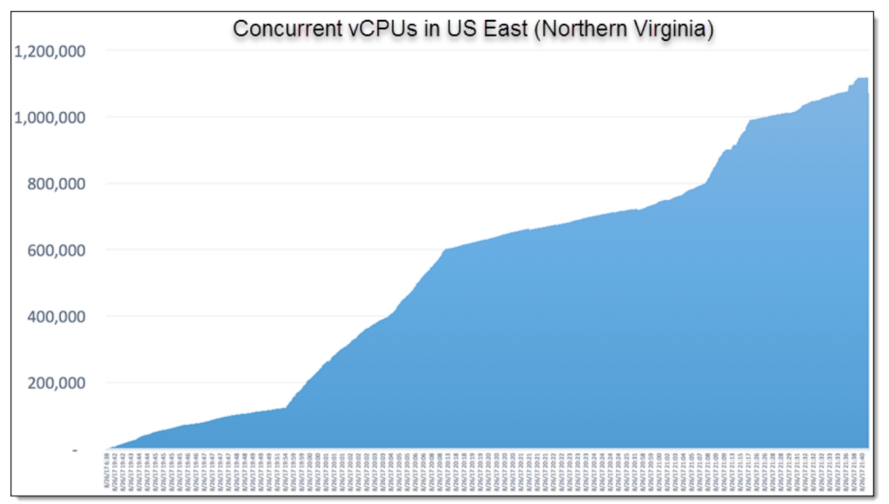
The combined resources of ADAPT and AWS reduce total processing time from 10 months to less than 1 month

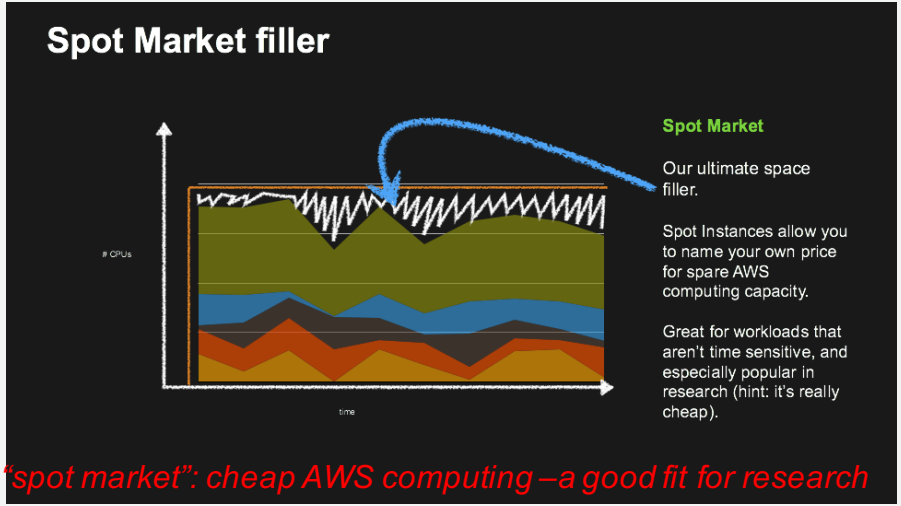
https://www.nas.nasa.gov/SC15/demos/demo31.html

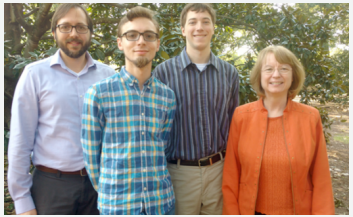
aws

Availability of resources: Natural Language Processing at Clemson University

550,000 cores using EC2 Spot Instances







"I am absolutely thrilled with the outcome of this experiment. The graduate students on the project […] used resources from AWS and Omnibond and developed a new software infrastructure to perform research at a scale and time-to-completion not possible with only campus resources." – Prof. Amy Apon, Co-Director of the Complex Systems, Analytics and Visualization Institute

https://aws.amazon.com/blogs/aws/natural-language-processing-at-clemson-university-1-1-million-vcpus-ec2-spot-instances/

Right-sized resources: Genomics processing on FPGA Accelerators

Children's Hospital of Philadelphia and Edico Genome Achieve Fastest-Ever Analysis of 1,000 Genomes

Orlando, Fla., Oct 19, 2017 – The Children's Hospital of Philadelphia (CHOP) and Edico Genome today set a new scientific world standard in rapidly processing whole human genomes into data files usable for researchers aiming to bring precision medicine into mainstream clinical practice. Utilizing Edico Genome's DRAGEN™ Genome Pipeline, deployed on 1,000 Amazon EC2 F1 instances on the Amazon Web Services (AWS) Cloud, 1,000 pediatric genomes were processed in two hours and 25 minutes.

GUINNESS WORLD RECORDS™
RECORD HOLDER

edico genome

… Available in "AWS App Store" (AWS Marketplace) for ~$24 / genome

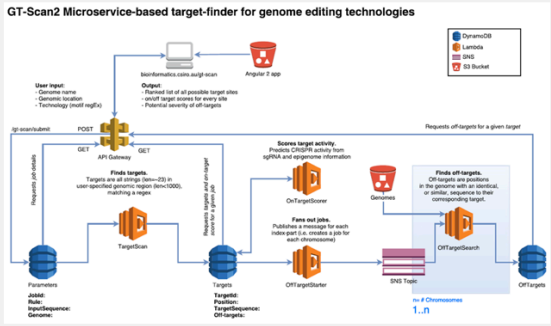
aws

Moving quickly with managed services

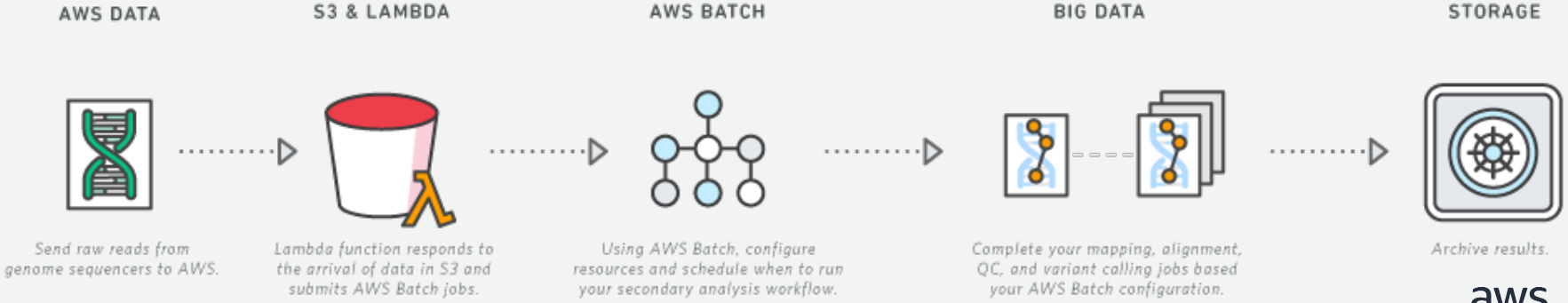
SageMaker: managed ML in notebooks

Serverless computing

CSIRO have built quickly scaling genomics analysis on AWS Lambda



DNA Sequencing using AWS container services



Moving quickly with managed services: CSIRO & CRISPR prediction

CSIRO is the federal government agency for scientific research in Australia

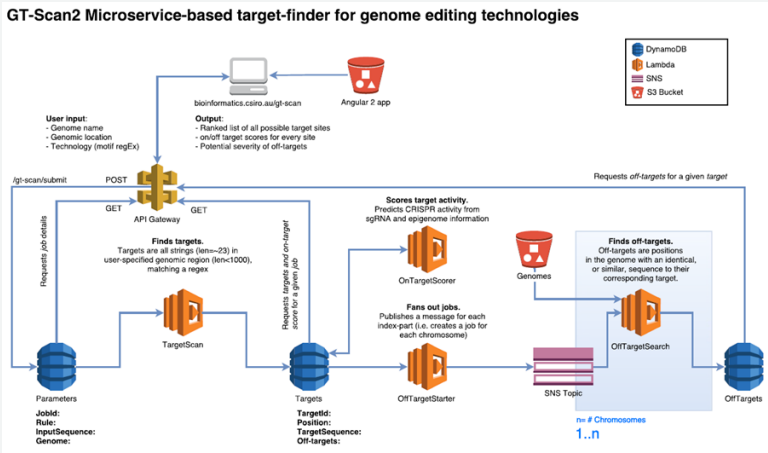
CSIRO **used AWS Lambda Serverless Computing functions to completely re-engineer a cluster HPC workload** to identify optimal gene editing sites for personalized treatment.

The job runtime varies from 1 second to 5 minutes, because the complexity of the targeted gene can vary dramatically. And the number of simultaneous jobs is unpredictable.

Server-based solutions can't handle the variability with quick turn-around – either you have lots of servers sitting idle, or you have to wait minutes for new servers to spin up.

With the Serverless microservices architecture, the GTScan-2 runtime is stable at a few minutes **per complete job**, no matter how many jobs (i.e. genetic samples) are sent to it.

Re-architecting the entire application took **only 3weeks**.



Key Strengths of AWS for Scientific Discoveries

Time to discovery

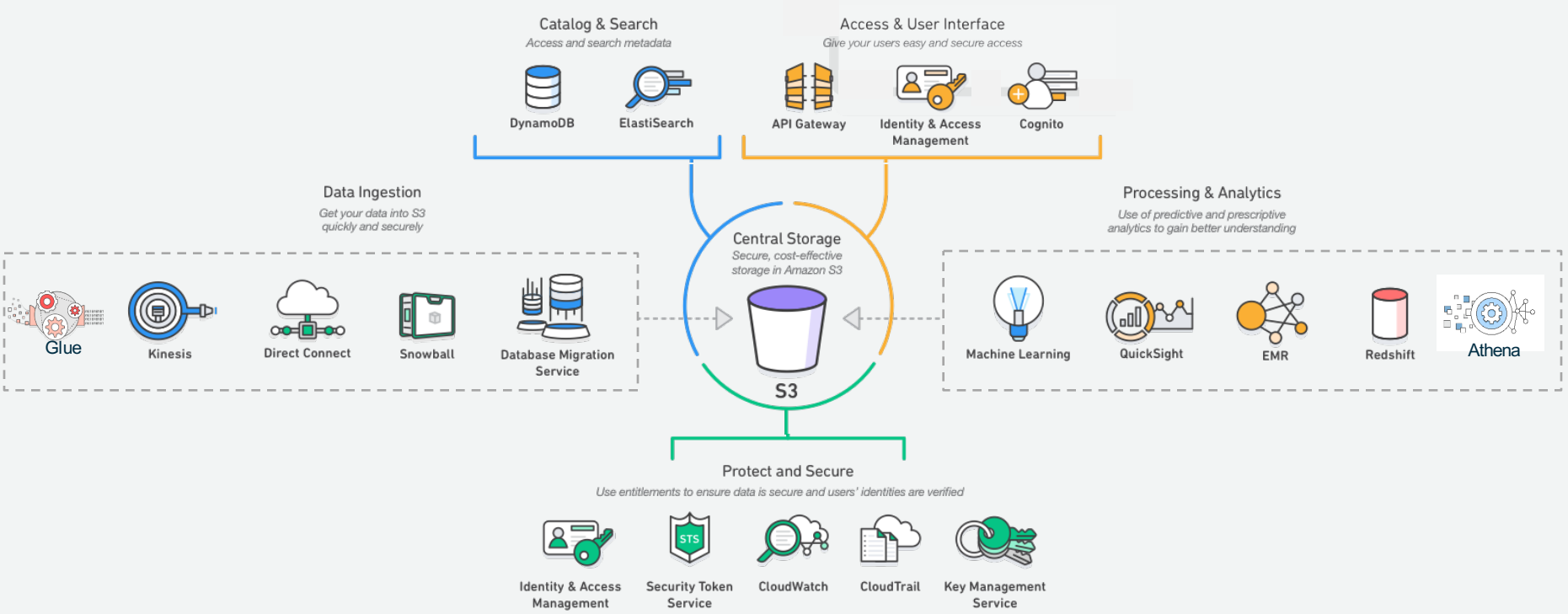
- Availability of resources, scalability, right-sizing
- Experiment fast
- Avoid undifferentiated work

Collaboration

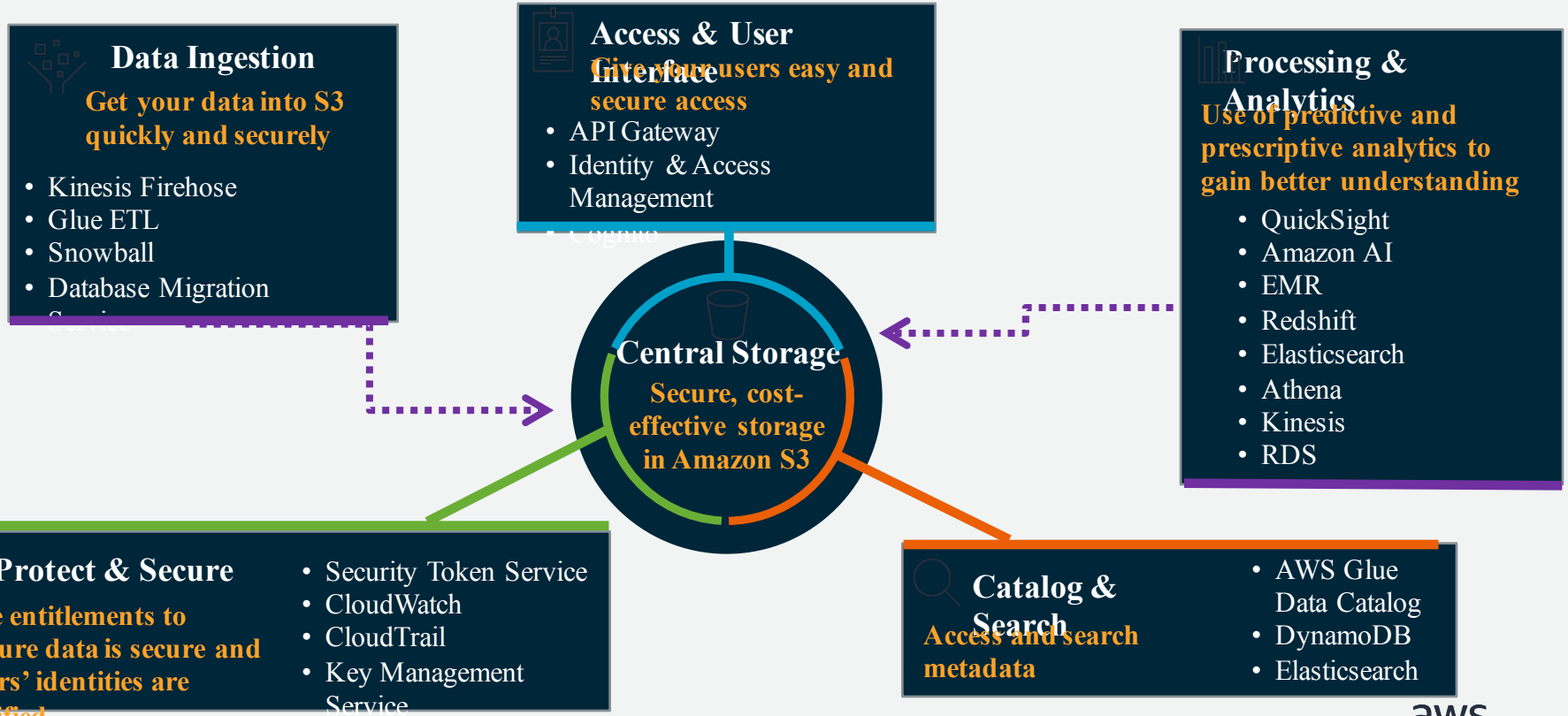
- Data lake model
- Security & compliance
- Sharing
- Infrastructure, ML, Analytics

aws

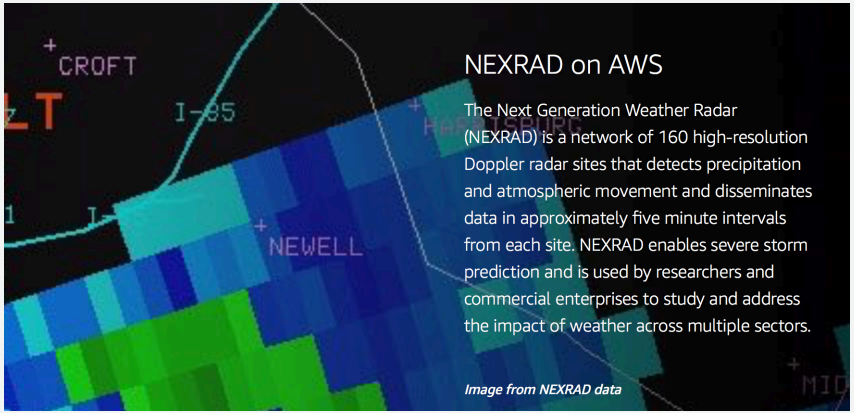
Collaborating on scientific data in the cloud

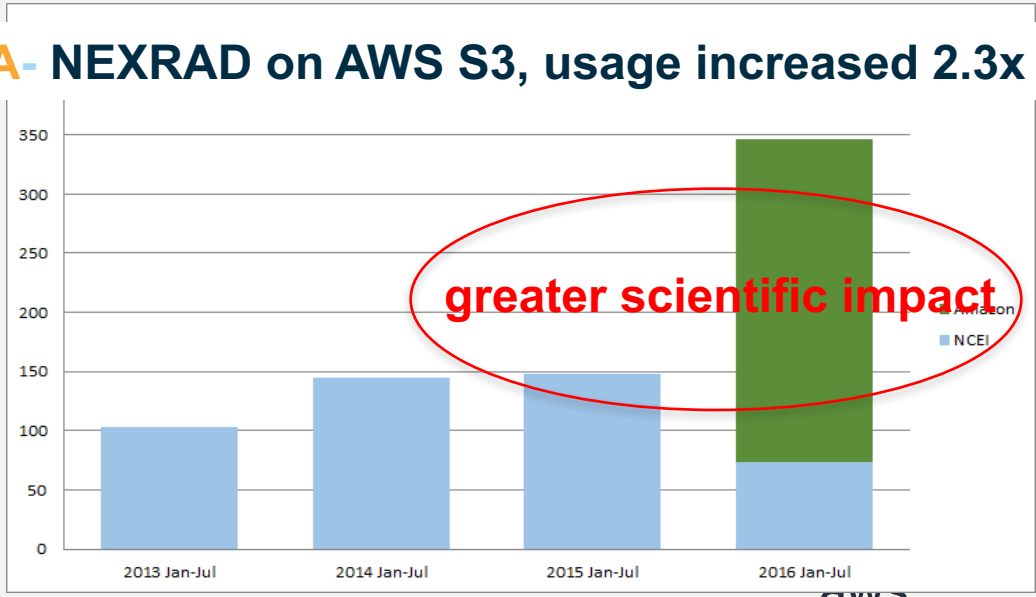


Getting Value Out of Your Data



Collaborating on scientific data in the cloud





Collaborating on scientific data in the clou

nature ecology & evolution

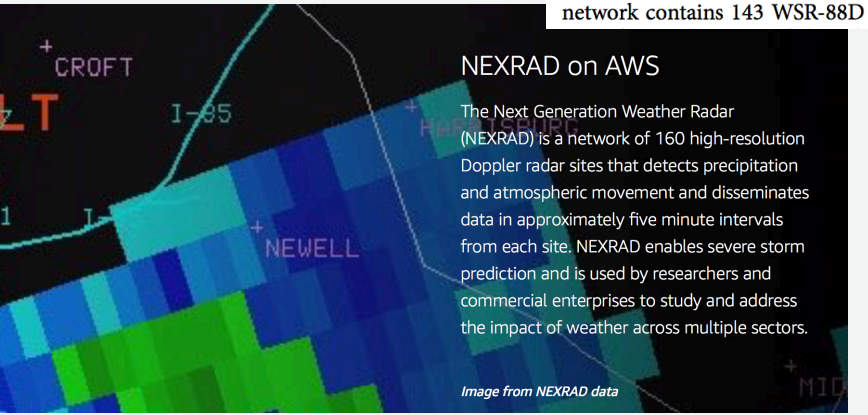
ARTICLES

https://doi.org/10.1038/s41559-018-0666-4

Seasonal abundance and survival of North America's migratory avifauna determined by weather radar

Adriaan M. Dokter[1*], Andrew Farnsworth[1], Daniel Fink[1], Viviana Ruiz-Gutierrez[1], Wesley M. Hochachka[1], Frank A. La Sorte[1], Orin J. Robinson[1], Kenneth V. Rosenberg[1,2] and Steve Kelling[1]

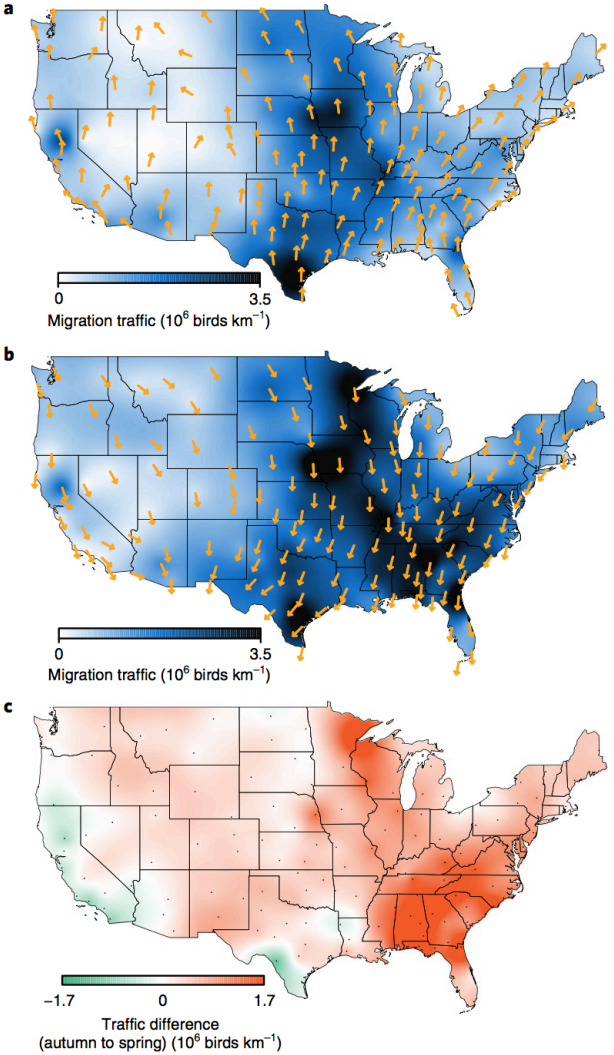
Recently, the National Oceanic and Atmospheric Administration and Amazon Web Services (AWS) Cloud made available one of the largest datasets describing animal movement ever compiled[20]: the Next Generation Weather Radar (NEXRAD) archive. The NEXRAD network contains 143 WSR-88D weather radars in the contiguous



NEXRAD on AWS

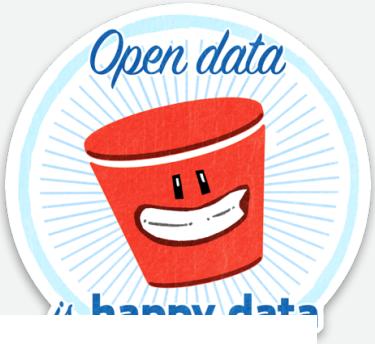
The Next Generation Weather Radar (NEXRAD) is a network of 160 high-resolution Doppler radar sites that detects precipitation and atmospheric movement and disseminates data in approximately five minute intervals from each site. NEXRAD enables severe storm prediction and is used by researchers and commercial enterprises to study and address the impact of weather across multiple sectors.

Image from NEXRAD data



NIH initiatives: National Cancer Institute

Funded projects to create collaborative environments on cloud



- Access and analyze 11,000 TCGA samples without having to download data
- Upload your own data for analysis

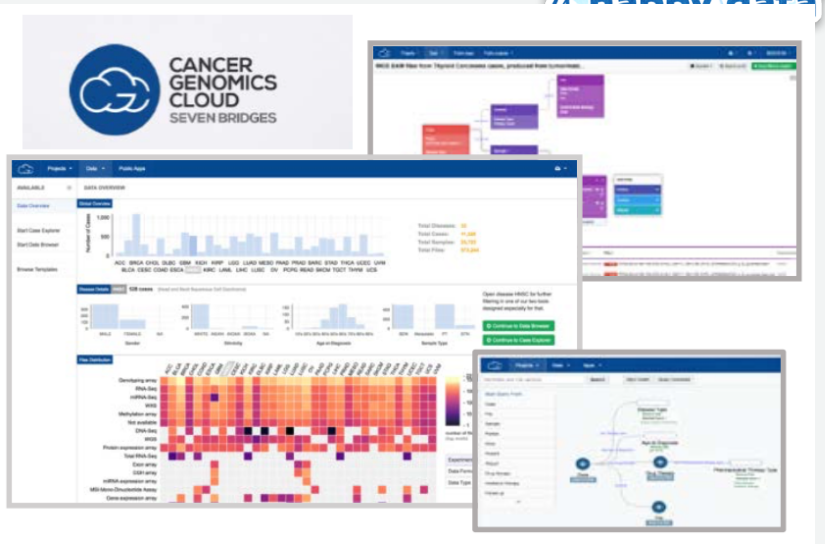
Data

- Perform large scale analysis using the elastic compute power of commercial cloud platforms

Compute

- dbGaP-authorized users can access controlled TCGA data
- Systems meet strict Federal security guidelines

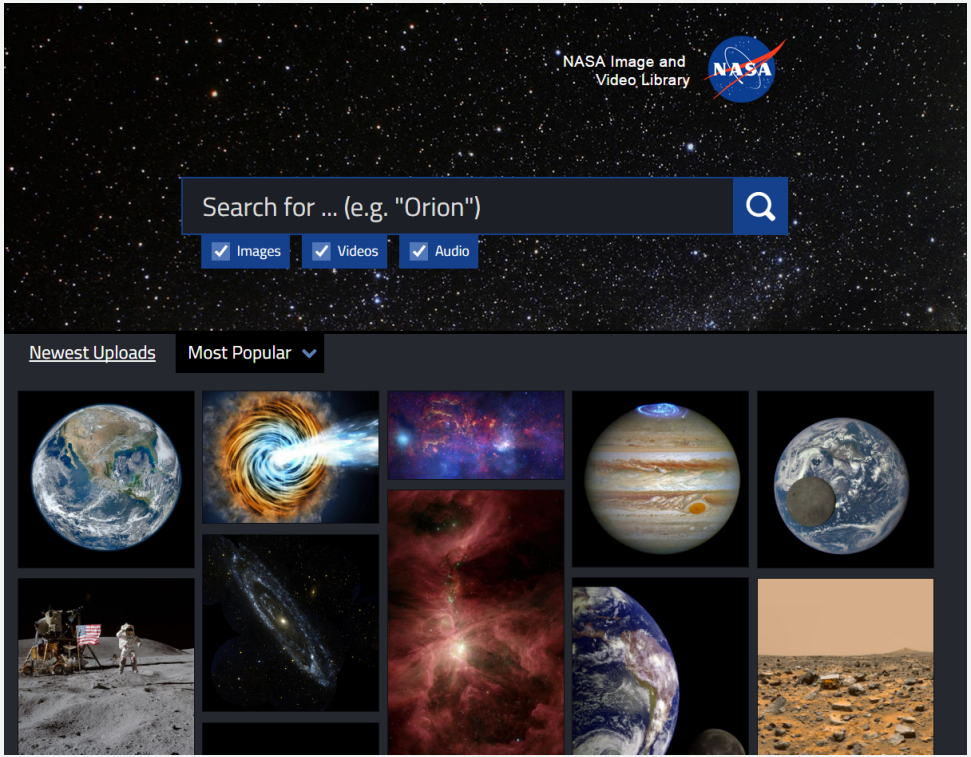
Security



http://www.cancergenomicscloud.org

aws

NASA Image and Video Library



- Easy Access to the Wonders of Space. Fully compliant with Section 508 of the Rehabilitation Act.

- Built-in Scalability. “On-demand scalability will be invaluable for events such as the solar eclipse that’s happening later this summer—both as we upload new media and as the public comes to view that content,” says Bryan Walls, Imagery Experts Deputy Program Manager at NASA.

- Good Use of Taxpayer Dollars. By building its Image and Video Library in the cloud, NASA avoided the costs associated with deploying and maintaining server and storage hardware in-house. Instead, the agency can simply pay for the AWS resources it uses at any given time.

https://aws.amazon.com/partners/success/nasa-image-library/

aws

U.K. Met Office Uses AWS to Deliver Tailored Meteorological Data

> "We are using the AWS Cloud to drive the mass-market availability of customizable weather information."

James Tomkins
Head of Enterprise IT Architecture
Met Office

The Met Office has been a widely respected national weather service in the United Kingdom for 160 years.

- Needed the means to send weather data to device users and third-party customers.
- Deployed Amazon ElastiCache to respond to peak demands.
- Attracted more than half a million users with its WeatherCloud app.
- Scaled data storage tenfold and reduced solution costs by 50 percent.
- Enabled innovation of big data services in a competitive landscape.

https://aws.amazon.com/solutions/case-studies/the-met-office/

https://aws.amazon.com/about-aws/whats-new/2017/08/uk-met-office-high-resolution-weather-forecast-data-is-now-on-aws/

Open Data on AWS

To stimulate innovation, AWS hosts a selection of datasets that anyone can access for free. Data in our public datasets is available for rapid access to our flexible and low-cost computing resources.

Life Science

- **TCGA & ICGC** (used at OICR)
- 1000 Genomes
- Genome in a Bottle
- Human Microbiome Project
- 3000 Rice Genome

Earth Science

- Landsat
- NEXRAD
- NASA NEX

Internet Science

- Common Crawl Corpus
- Google Books Ngrams
- Multimedia Commons

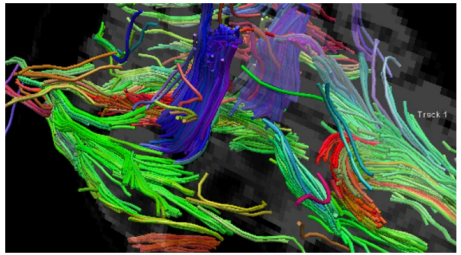
https://aws.amazon.com/public-datasets/

aws

Open Data on AWS

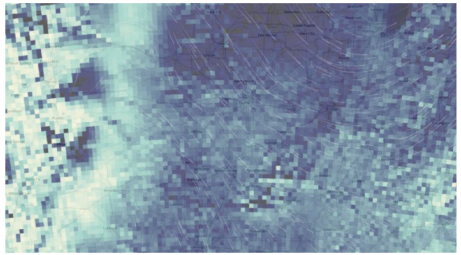
Visit **Earth on AWS** to learn about building planetary-scale applications in the cloud with open geospatial data.

All public data from the **Hubble Space Telescope**'s active instruments are available for large-scale analysis on Amazon S3.



The **Allen Institute for Brain Science** and the **University of Washington** provided students with 35TB of data with Amazon S3.

You can query billions of **OpenStreetMap** features with Amazon Athena without needing to download data or set up a server.



The **National Renewable Energy Laboratory (NREL)** makes a 500 TB open weather model dataset available to the world on Amazon S3.

```
      sv.referenceallele ,
      sv.alternateallele ,
      sv.genotype0 ,
      sv.genotype1
FROM demo.samplevariants sv
CROSS JOIN
  (SELECT count(1) AS numsamples
   FROM
     (SELECT DISTINCT sampleid
      FROM demo.samplevariants
      WHERE sampleid LIKE 'NA12%'))
JOIN demo.clinvar cv ON sv.chromosome = cv.chromosome
AND sv.startposition = cv.startposition - 1
AND sv.endposition = cv.endposition
AND sv.referenceallele = cv.referenceallele
AND sv.alternateallele = cv.alternateallele
WHERE assembly='GRCh37'
  AND cv.clinicalsignificance LIKE '%response%'
```

Learn how to prepare **1000 Genomes** data for fast interactive analysis using Amazon Athena.

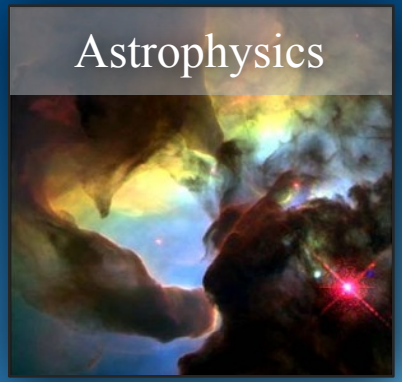
https://aws.amazon.com/opendata/

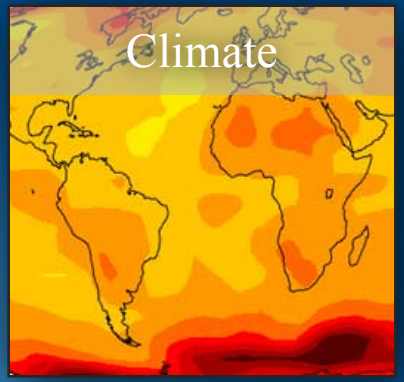
aws

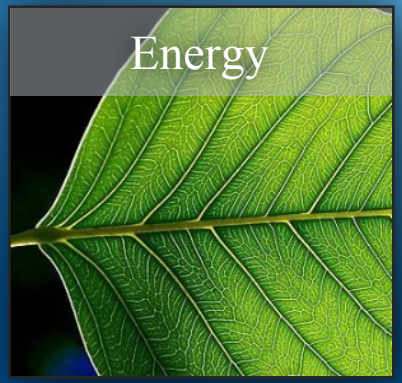
Making Fast & Reliable HPC Possible

Akanksha Balani
Intel® Software

HPC Enables Insight and Fuels Innovation

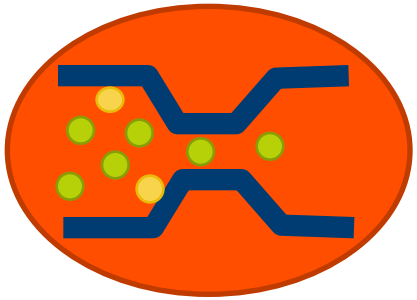






The growing challenge in hpc

System Bottlenecks "The Walls"



Memory | I/O | Storage Energy Efficient Performance Space | Resiliency | Unoptimized Software

Divergent Workloads

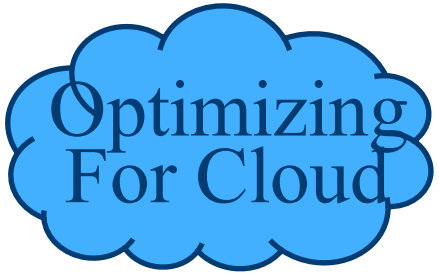
Machine learning

hpc  Big Data

visualization

Resources Split Among Modeling and Simulation | Big Data Analytics | Machine Learning | Visualization

Barriers to Extending Usage



Democratization at Every Scale | Cloud Access | Exploration of New Parallel Programming Models

Intel accelerating high performance computing

ecosystems

platforms

Intel Parallel Computing Centers
Intel® Dev CLoud
Intel® DL Studio
Intel® OpenVino ToolKit
Movidius Fathom
intel Saffron

Frameworks

Spark MLlib BigDL
neon
TensorFlow
Microsoft CNTK
mxnet
torch
Caffe
Caffe2
Chainer
theano

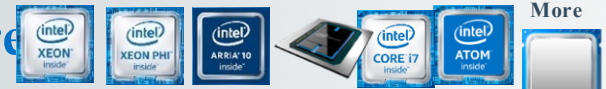
libraries HPC

python Intel Python Distribution
Intel® Data Analytics Acceleration Library (DAAL)
Intel® Nervana™ Graph*
Intel® Math Kernel Library (MKL, MKL-DNN)

Software Tools

PARALLEL STUDIO XE
Compilers – C/C++/Fortran
Analysis Tools – Vtune, Advisor, Inspector
Libraries – Math Kernel Library, TBB, IPP, DAAL, MPI Library
Cluster Checker and Trace analysis tools

hardware

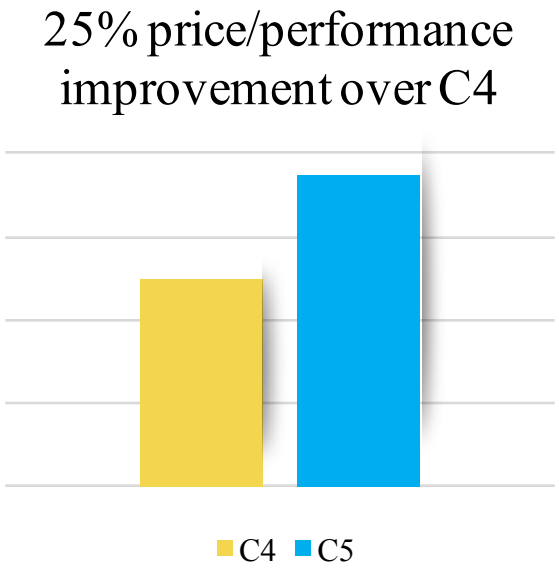


Compute
More
Memory & Storage
Networking

*Future

aws intel

C5: Compute-optimized instances based on Intel Skylake



- **Based on 3.0 GHz Intel Xeon Scalable Processors (Skylake)**
- **Up to 72 vCPUs and 144 GiB of memory (2:1 Memory:vCPU ratio)**
- **25 Gbps NW bandwidth**
- **Support for Intel AVX-512**

NETFLIX

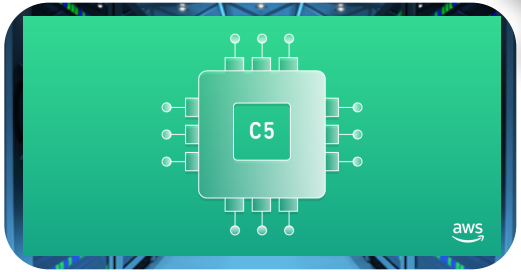
"We saw significant performance improvement on Amazon EC2 C5, with up to a 140% performance improvement in industry standard CPU benchmarks over C4."

GRAIL

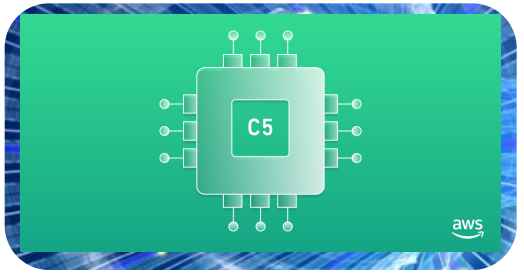
"We are eager to migrate onto the AVX-512 enabled c5.18xlarge instance size… . We expect to decrease the processing time of some of our key workloads by more than 30%."

Performance Drivers for HPC applications

Compute



Bandwidth



SW Optimizations

Intel software suite for HPC & Compute

Edge to DC to Cloud

Manuf., Retail, Drones, Robots…

Smart Cities, Auto. Driving, Gaming…

Fast, Dense, High Quality Transcoding

Technical & Enterprise compute, HPC, AI

Take advantage of deep system-wide insight & analysis for system & embedded apps

Optimization Tools , SDKs

Create solutions using Computer Vision – OpenVino Toolkit, Deep Learning, Graphics, Libraries, Media, OpenCL™, & more

Build highly optimized media infrastructure, solutions, & applications

Improve performance, scalability, & reliability for applications and frameworks - Computing and ML/DL

Big DL

Intel® Distribution of Python

Intel® DAAL

aws intel

Intel® Parallel Studio XE - Create Faster Code…Faste

Composer Edition

BUILD
Compilers & Libraries

- C / C++ Compiler — Optimizing Compiler
- Intel® MKL — Fast Math Kernel Library
- Fortran Compiler — Optimizing Compiler
- Intel® IPP — Image, Signal & Data Processing
- Intel® TBB — C++ Threading Library
- Intel® DAAL — Data Analytics Library
- Intel® Distribution for Python* — High Performance Scripting

…sional Edition

ANALYZE
Analysis Tools

- Intel® VTune™ Amplifier — Performance Profiler
- Intel® Inspector — Memory & Thread Debugger
- Intel® Advisor — Vectorization Optimization & Thread Prototyping

Clu… Edition

SCALE
Cluster Tools

- Intel® MPI Library — Message Passing Interface Library
- Intel® Trace Analyzer & Collector — MPI Tuning & Analysis
- Intel® Cluster Checker — Cluster Diagnostic Expert System

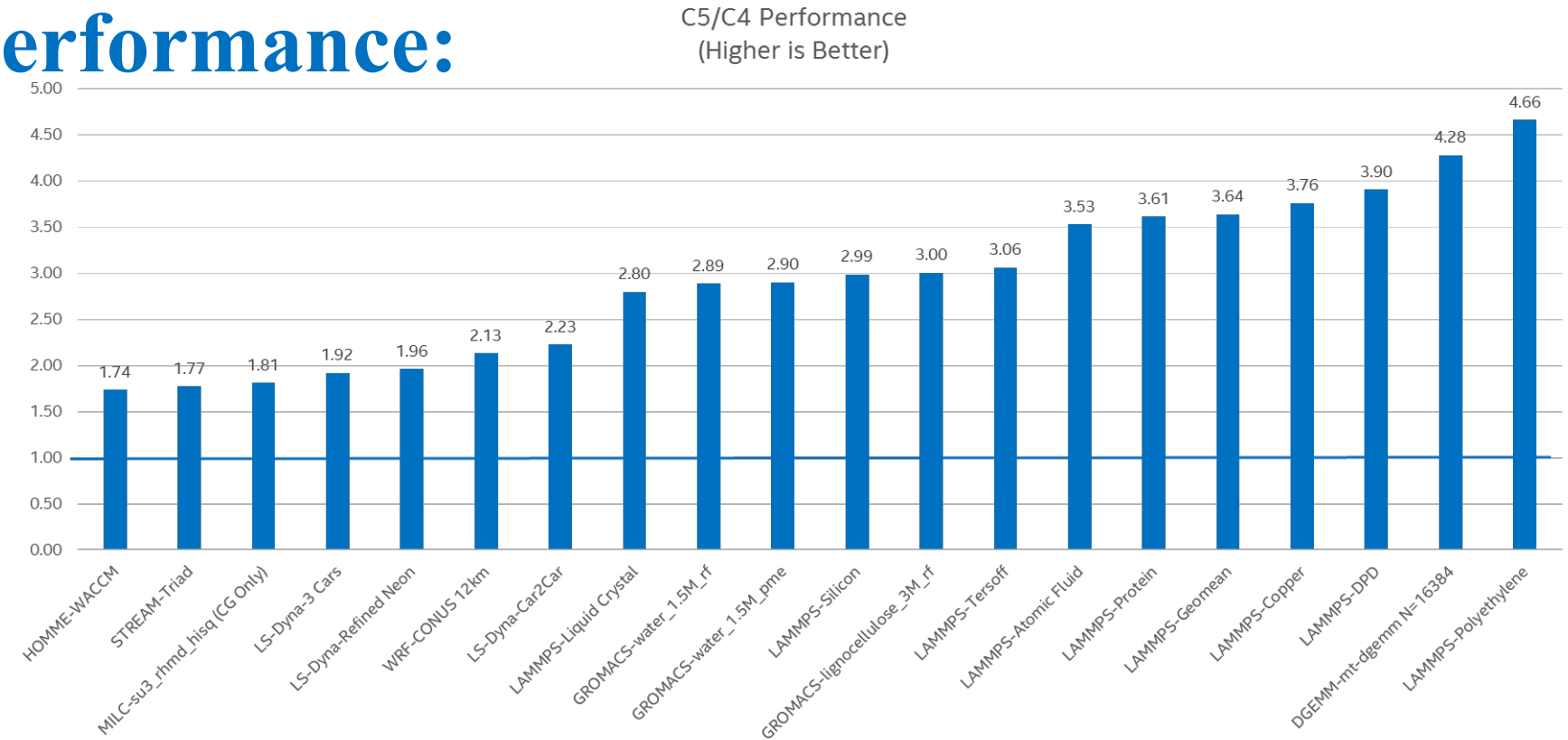
Intel® Architecture Platforms

Operating System: Windows*, Linux*, MacOS[1]*

More Power for Your Code - software.intel.com/intel-parallel-studio-xe

aws | intel

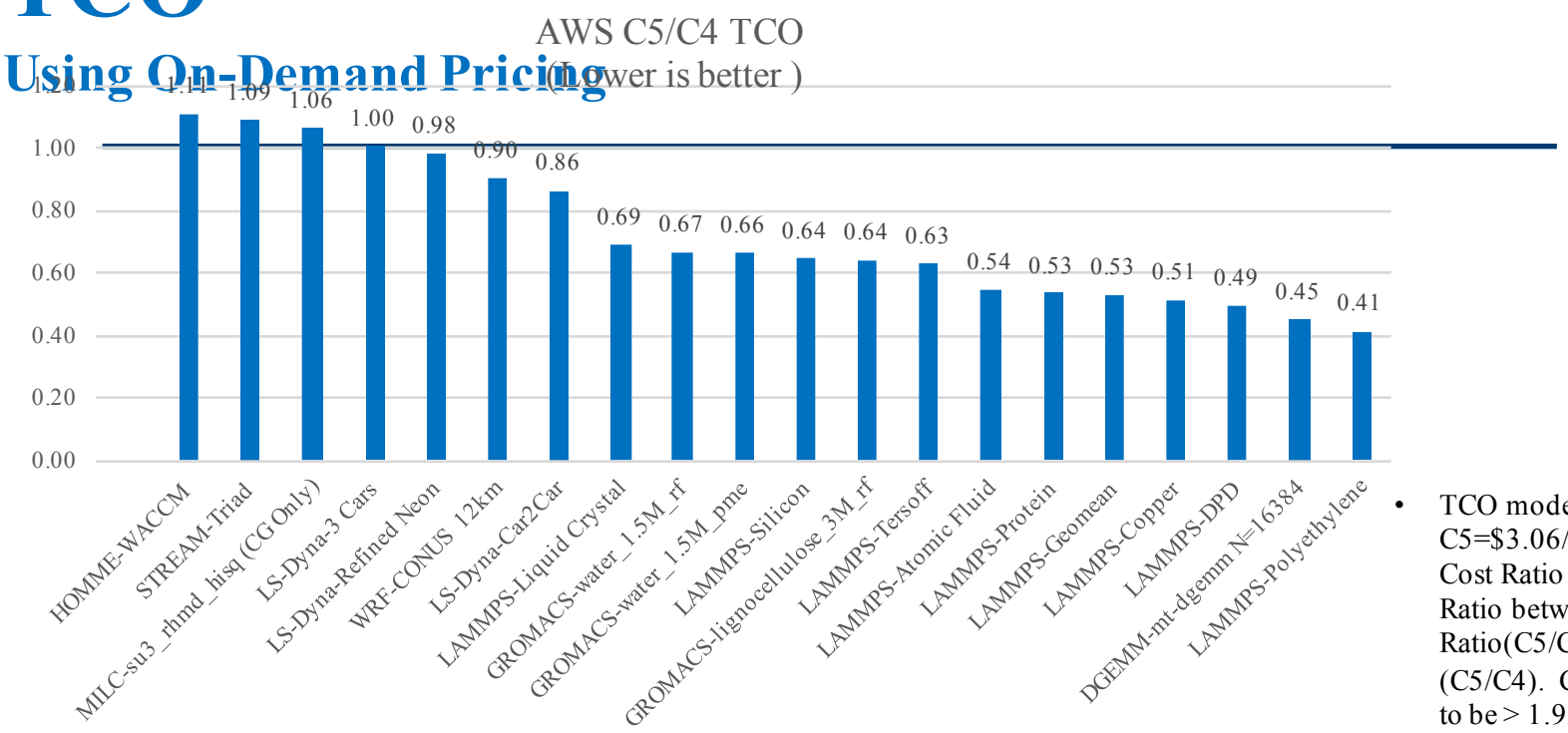
Application-Workloads - Performance:



Testing conducted on HPC applications and workloads comparing AWS C4.8x vs C5.18x instances. Testing by Intel. For complete testing configuration details, **see the Configuration Details section (slide 5)**

Application-Workloads – TCO

Using On-Demand Pricing



Higher TCO with C5

Reduced TCO with C5

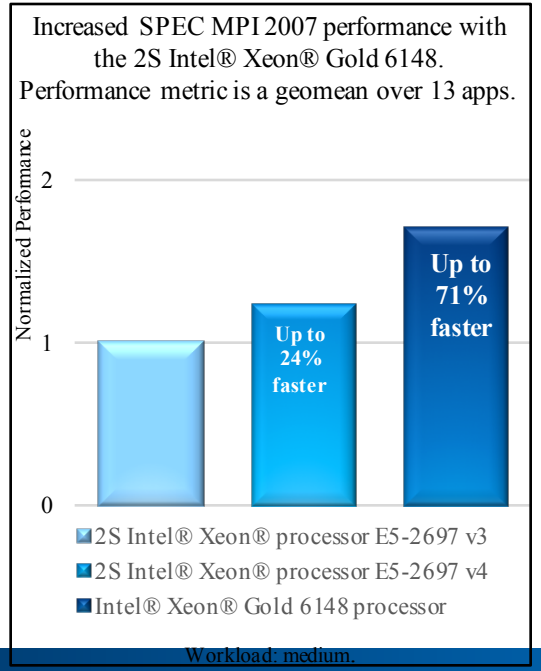
- TCO model: Given Cost: C5=$3.06/Hr, C4=$1.59/Hr. Cost Ratio C5/C4=1.92.TCO Ratio between C5/C4= Cost Ratio(C5/C4) / Perf Ratio (C5/C4). C5/C4 perf ratio needs to be > 1.92 to be cost efficient

Testing conducted on HPC applications and workloads comparing AWS C4.8x vs C5.18x instances. Testing by Intel. For complete testing configuration details, **see the Configuration Details section (slide 5)**.

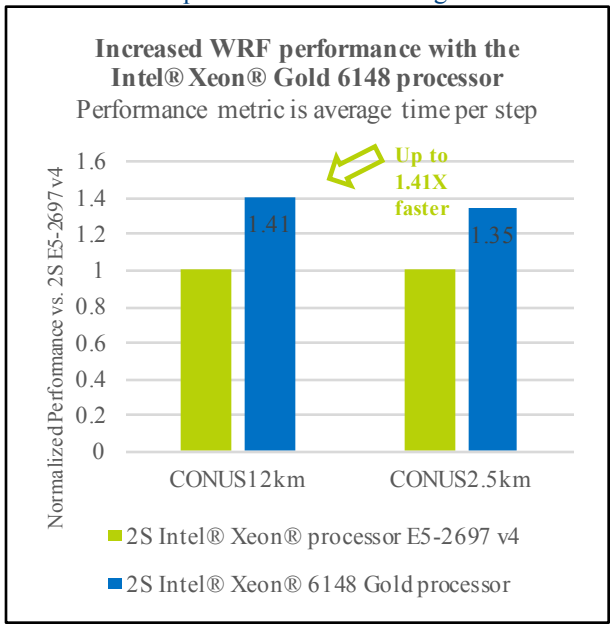
Intel® Xeon® processor scalable family

Scalable performance for widest variety of HPC workloads

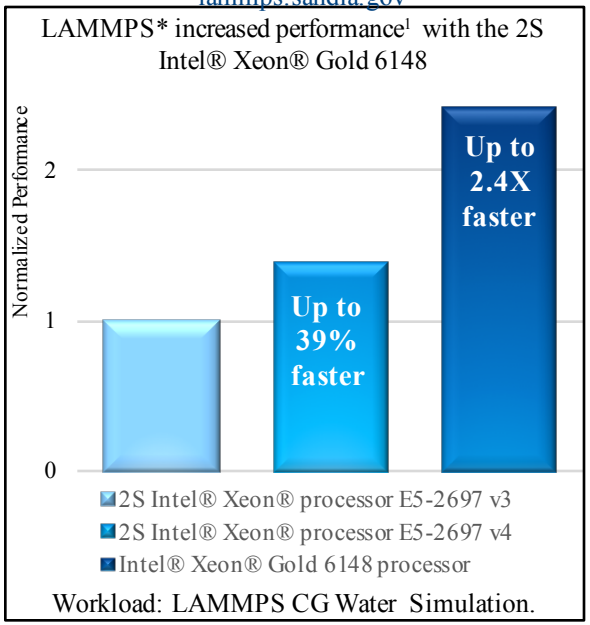
www.spec.org/mpi2007



https://www.wrf-model.org/



lammps.sandia.gov



Find out more

learn

More information at www.intel.com/hpc or www.intel.com/software/products and AWS & Intel Technology Forums

explore

New instances based on Intel Xeon Scalable on AWS

engage

Contact your Intel representative for help and POC opportunities #booth


aws
intel

Legal Disclaimer & Optimization Notice

INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804