

米国・欧州

各国のトップエンドHPC開発計画と 気象気候研究との関わり

井上孝洋@JAMSTEC

2022/05/16

第1回計算科学研究連絡会(気象・気候分野)

自己紹介・Disclaimer

- 経歴
 - 1998年頃からなんらかの形でHPCに関わってる(が、研究者ではない)
 - 東大CCSR→(民間シンクタンク)→ RIST・理研(出向)・JAMSTEC(招聘)
 - 現職はRESTEC/JAMSTEC招聘
 - 共生・革新・創生・統合・先端(予定)各プログラムになんらかの形で参加
 - HPCI戦略プログラムでは取りまとめる側(MEXT委託業務)
 - 若かりし頃はHPC向けチューニングもやっていたが、近年は動向調査が主。
- 今日のお話
 - 概要をざっと。詳細は端折り気味で。
 - オープンになっている情報ベース。裏情報や特ダネがあるわけでは無い。
 - 計算能力の話だけ。
 - ストレージの話も大事ですが、今日はしません。
 - ソフトウェアの話はこの後の講演に任せます。
- 主な情報源
 - 19th Workshop on high performance computing in meteorology (HPCWS2021)
 - HPC USER Forum September 2021, March 2022
 - HPCWire(英語、日本語)
 - 各機関のWebサイト
- 一部の図・写真・内容の引用元はリンクとして入れてあります。

アジェンダ

歴史とトレンド

米国のHPC

まとめ

欧州のHPC

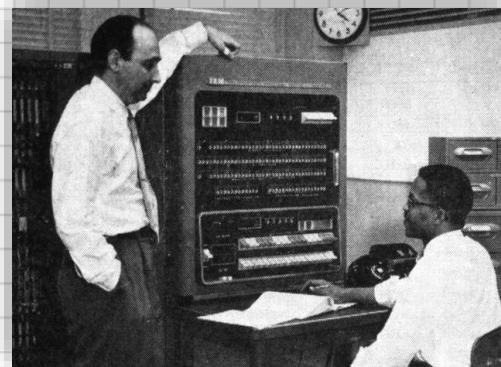
歴史とトレンド

HPCと気象気候

- リチャードソンの夢 (1922)
- Charneyの数値予報 (1950)
- JNWPU, IBM701 (1955)
- 気象庁, IBM704 (1959)
- 「気象庁は、予報に使える大型コンピュータを前から要求していたが、1957年に認められ(おそらく1958年度予算として)、1959年初め、IBM 704を設置した。3月12日、数値予報業務を開始。7月、台風5号の進路予想に活躍した。」
(「新HPCの歩み -1959年(a)-」小柳義男)
- 「IBM704は日本政府が行政用に導入した初めてのコンピュータで、導入当時は大きな話題となりました。ただし、その性能は今日のパーソナルコンピュータにも遠く及ばないため、当初の数値予報結果は現場の予報官の使用には耐えず、予報官の信頼を得るまでにはかなりの年月を必要としました。」
(「数値予報の歴史」気象庁)



<https://www.jma.go.jp/jma/kishou/known/whitep/1-3-2.html>



https://celebrating200years.noaa.gov/foundations/numerical_wx_pred/701computer.html



<https://www.jma.go.jp/jma/kishou/known/whitep/1-3-2.html>

HPCなかりせば、気象気候研究は不可能と言ってもいい。
HPC製品も以前は主たる顧客は気象気候研究が多かった(気がする)。

気象・気候分野とHPCベンダーが一同に介する機会

最先端並列計算機における次世代気候モデル開発に関わる国際ワークショップ (1999~2012)

[シンポジウム] 306:504 (並列計算技術:並列計算機)

最先端並列計算機における次世代気候モデル開発に係わるワークショップ報告*

山岸 米二郎*¹ 室井 ちあし*² 保坂 征宏*³

1. はじめに
 標記ワークショップが、東京大学気候システム研究センター (CCSR) 所長住教授と NCAR の全球大気物理学と気候研究部長 Blackmon 両氏をコンピーナーとして、1999年3月1日から3日間ハワイ大学東西センターで実施された。この会議は科学技術振興調整費総合研究「高精度の地球変動予測のための並列ソフトウェア開発に関する研究」の大気・海洋分野分科会の活動として、開催された (財団法人高度情報科学技術研究機構 (以後 RIST と略称) 主催、科学技術庁後援)。参加者は延べ83名 (日本からは38名、その他は米国) で、最先端並列計算機、次世代気候モデル、計算機メーカーの現状と将来、討論の4つのセッションで実施され、29の発表 (日本から12) があった。
 我が国の気象界でも分散主記憶型の並列計算機がかなり利用可能になっているので、会員の参考に供するために、並列計算を重点に並列計算機、気候研究、モデル開発、討論の項目に分けてワークショップの内容を報告する。最後に3人の感想も付した。なおこれまで並列計算機の記事が少なかったので、付録に簡単な用語解説をつけた。

2. 並列計算機


ド) あり、1Nあたり8PE (プロセッサエレメント)、1PEのピークスピードが8Gflops (1ギガフロップス=毎秒10⁹浮動小数点演算)、記憶容量は1N当たり16G (ギガ) バイトの共有メモリーという構成である。全体で記憶容量10T (テラ=10¹²) バイト、ピークスピード40Tflopsという文字通りのベクトル型超並列計算機である。この振興調整費研究はESの活用を目指した並列ソフト開発研究である。

2002年春の運用開始を目指して着々進行中の計画の具体的紹介なので、米国側参加者から強い関心が示された。関心の1つは勿論計算機ハード関連で、使用電力量など細部についてまで質問があった。関心のもう1つは膨大な計算結果の処理である。データ処理は基本OSとか画像表示技術等に加え、シミュレータからユーザまでのデータ転送、ユーザが個別に使用するハード等、シミュレータ計画だけでは解決できない事項にも関係するので直ちに解決案が示され得るわけではない。ユーザにとっては計算機の性能向上により高分解能モデルの運用が可能になったとき、データの洪水に溺れずにそこからどのように効率的に有効な情報を得るかという挑戦となる。

Buzbee (Buzbee Enterprises) は今後の HPC (High Performance Computer) のアーキテクチャが、DSM

「天気」1999.11

iCAS/NCAR



September 2022
Stresa, Italy

International Computing for the Atmospheric Sciences Symposium

CISL/NCAR

HPCWS/ECMWF



20th ECMWF workshop
High Performance Computing in Meteorology

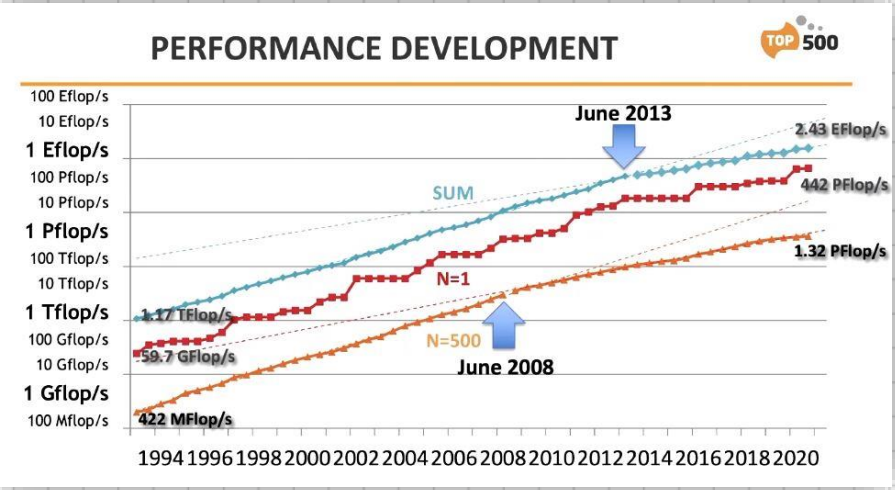
Bologna, Italy
September 2023

他に (特に日本/アジアで) あれば教えて下さい。

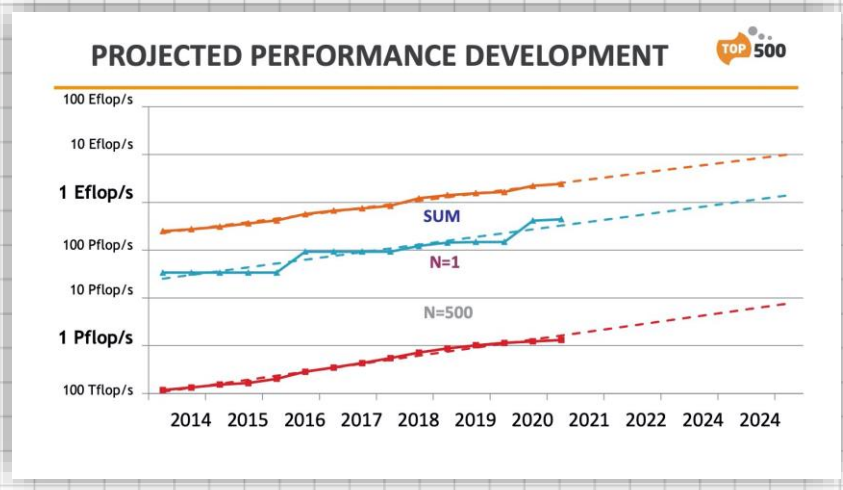
https://events.ecmwf.int/event/169/contributions/2785/attachments/1468/2651/HPC-WS_Weger.pdf

HPCのトレンド

もうすぐExa

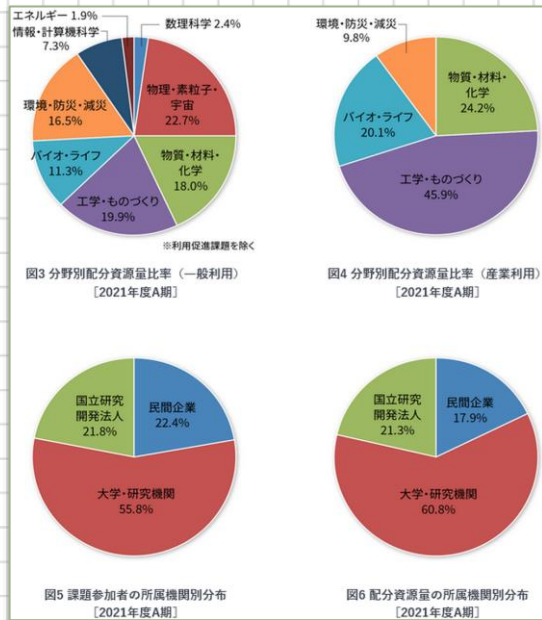


https://twitter.com/HPC_Guru/status/1328461458124931074



<https://twitter.com/top500supercomp/status/1328501542442340352>

物質科学やその他分野に広がる。
何より、AI/ML。



<https://www.r-ccs.riken.jp/fugaku/fugaku-annual-reports/2020/3/2>

“Extreme Heterogeneity”

HPCWS2021の最初のキーノートのテーマ

OAK RIDGE
National Laboratory

Preparing for Extreme Heterogeneity in High Performance Computing

Jeffrey S. Vetter
With many contributions from ACSR Section and Colleagues

19th Workshop on HPC in Meteorology
21 Sep 2021
ECMWF (Virtual)

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF
ENERGY

Future
Technologies
Group

<https://www.ornl.gov/section/advanced-computing-systems-research> (<https://j.mp/acsrns>)
vetter@computer.org

Highlights

Recent trends in computing paint an ambiguous future for architectures

- Power constraints initially drove architectural changes
- Now, vendors are forced to use 2-3 foundries if they want access to leading-edge CMOS production
- Forces vendors to add value with domain specific architectures by specializing processors, node design, memory systems, I/O
- Explosion of new architectures
- Devices: GPUs, FPGAs, DSPs, SoCs
- Deployment: HPC, AI, Edge, Cloud
- OpenHW: RISC-V

Entering an era of Extreme Heterogeneity



As a result, applications and software systems are all reaching a state of crisis

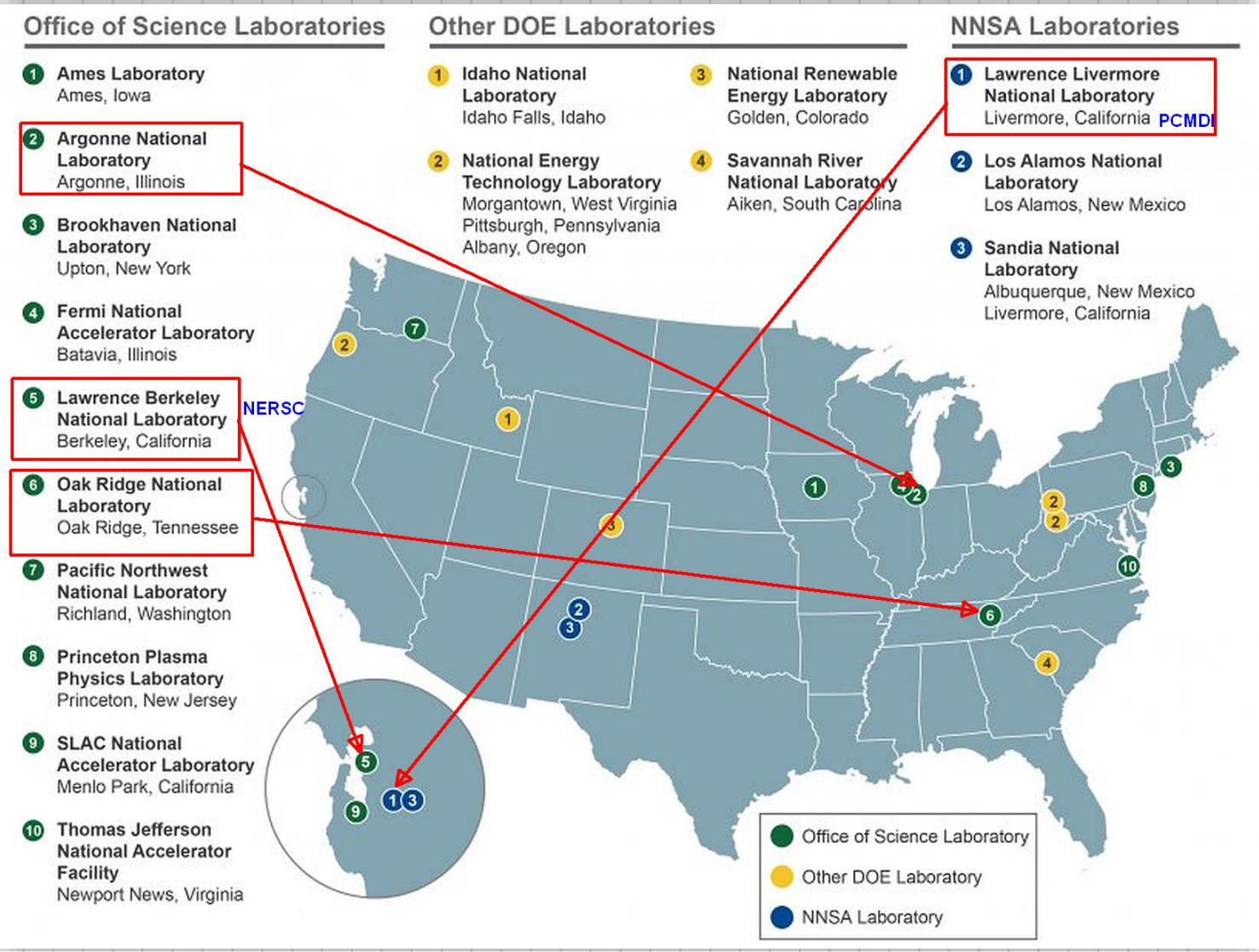
- Proliferation of diverse and often immature programming ecosystems
- In fact, programming and operating systems need major investment to address current and future architectural changes
- Applications will not be *functionally* or performance portable across architectures
- Additionally, procurements, acceptance testing, and operations of today's new platforms depend on performance prediction and benchmarking.
- **Complexity is our main challenge**
- **This is a crisis!**

Programming systems must provide performance portability (beyond functional portability)!!

- Ultimately, we should strive for *'Write once, perform satisfactorily anywhere'*
- Descriptive models of parallelism and data movement
- Introspective runtime systems
- Layered, modular, open-source approaches required
- Performance prediction tools for design, procurement, and operations
- **Examples**
 - ECP investments in LLVM
 - FORTRAN with GPU offloading
 - Introspective Runtime Systems
 - Programming FPGAs
 - Without Verilog

米国のHPC

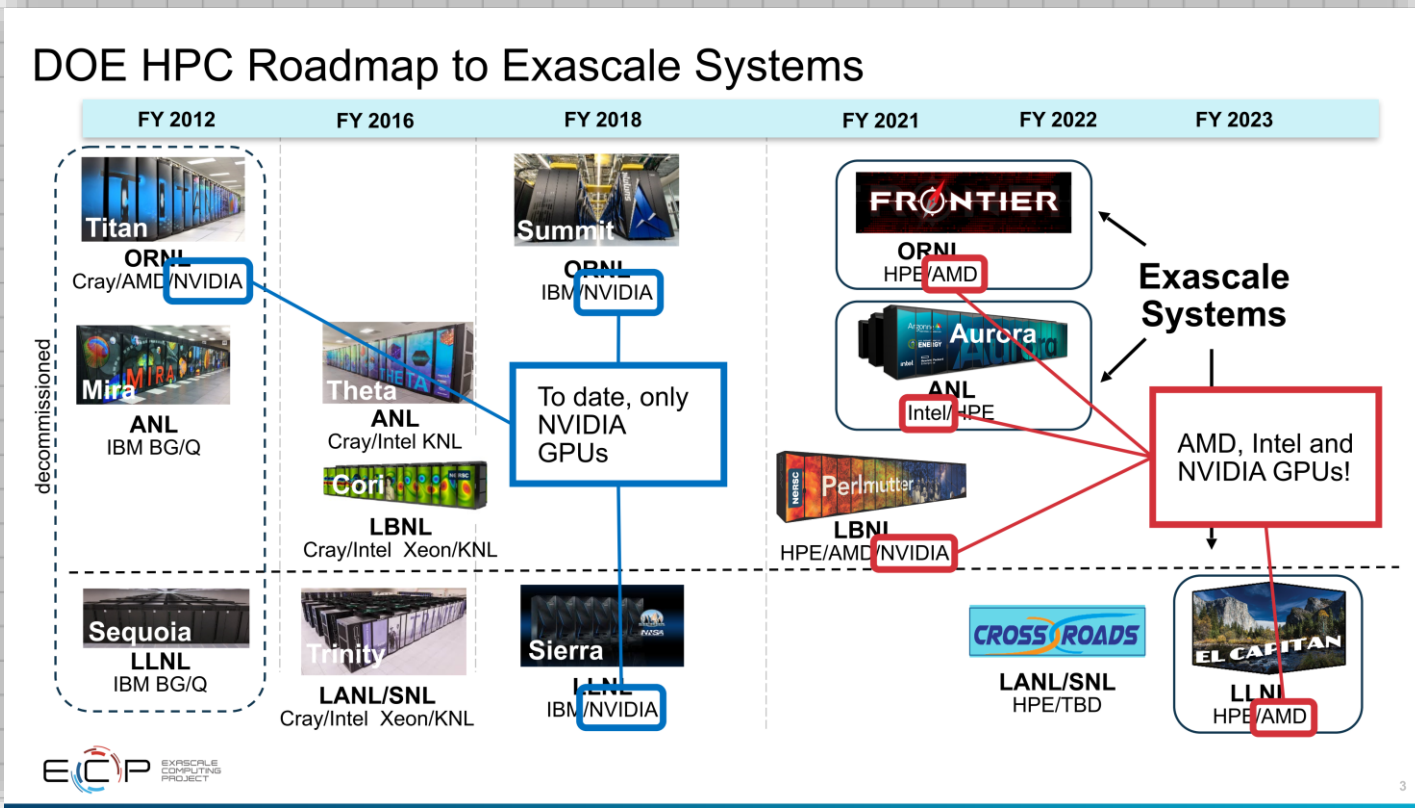
米国の主なHPCサイト (≒ DOEの研究所)



<https://www.energy.gov/doe-national-laboratories> より作成

The DOE-led Exascale Computing Initiative (ECI)

- DoEのOffice of Science (SC) とNational Nuclear Security Administration (NNSA)で2016年設立。2020年代半ばまでにExascale Computingを実現することを目的。
- アプリ開発: SCのBiological and Environmental Research program と Basic Energy Sciences program およびNNSA
- システム調達: ALCF-3 (“Aurora”), OLCF-5 (“Frontier”), ASC ATS-4 (“El Capitan”)
- Exascale Computing Project : 研究・開発・実装を主導



https://www.hpcuserforum.com/wp-content/uploads/2021/09/Exascale-Computing-Project-Update_ORNL_D.Kothe_Sept-2021-HPC-UF.pdf

ORNL:Frontier, ANL:Aurora



- <https://www.olcf.ornl.gov/frontier>
- Peak: > 1.5EFlops
- HPE Cray's new EX architecture and Slingshot interconnect
- CPU:HPC and AI Optimized 3rd Gen AMD EPYC CPU
- GPU:Purpose Built AMD Instinct 250X GPUs
- 2022年中(後半?)に運用開始

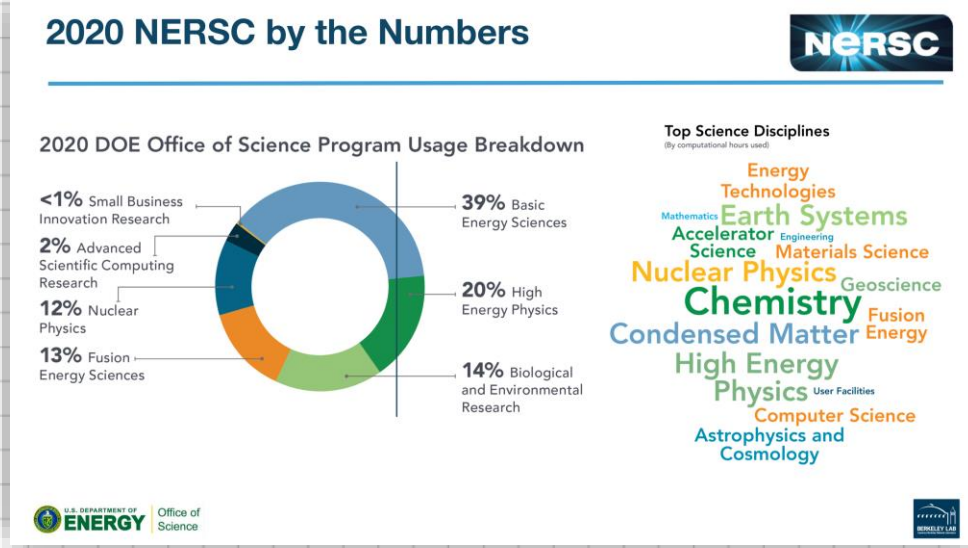
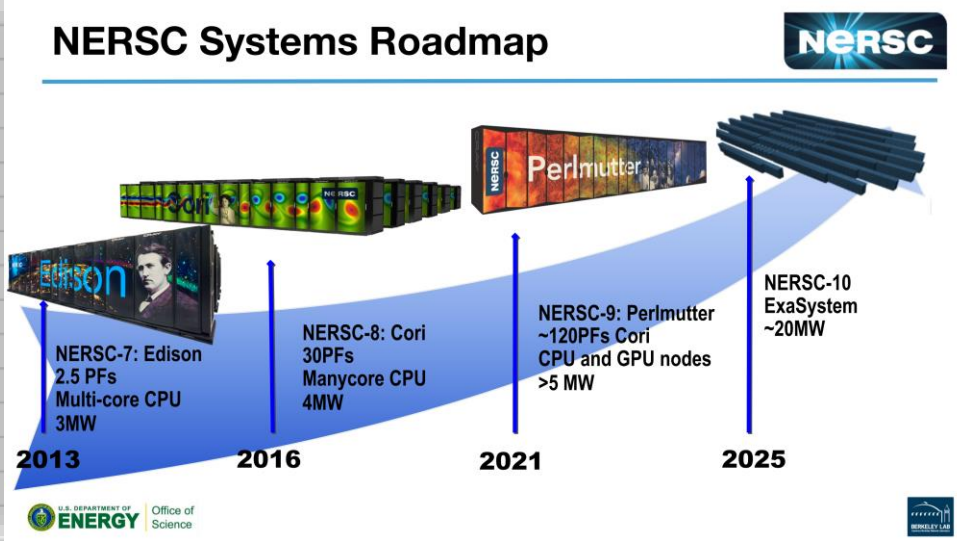


- <https://www.alcf.anl.gov/aurora>
- Peak Performance ≥ 2 Exaflop DP
- Arch: HPE's Cray EX architecture with Slingshot networking.
- CPU: two Intel Shappire Rapids
- GPU: six Ponte Vecchio
- OneAPI as a new model for programming the nodes across the entire system
- 2022年中にFull Delivery

Energy Exascale Earth System Model (E3SM) の研究開発、特に“A Baseline For Global Weather and Climate Simulations At 1KM Resolution”などに利用される予定。

NERSC:Perlmutter

- National Energy Research Scientific Computing Center
- LBNLの1ディビジョン
- NERSC is the principal provider of high performance computing and data resources and services to Office of Science programs — Advanced Scientific Computing Research, Basic Energy Sciences, Biological and Environmental Research, Fusion Energy Sciences, High Energy Physics, and Nuclear Physics.
- 学術利用のためのHPC



Perlmutter

- GPU-accelerated (GPU/CPU) and CPU-only nodes
- HPE Cray Slingshot high-performance network
- All-Flash filesystem
- Application readiness program (NESAP)
- HPE Cray System ~120PFs

Phase I: Arrived spring of 2021

- 1,536 GPU-accelerated nodes
- 1 AMD "Milan" CPU + 4 NVIDIA A100 GPUs per node
- 256 GB CPU memory and 40 GB GPU high BW memory
- 35 PB FLASH scratch file system
- User access and system management nodes

Phase II Addition: Arrives This Winter

- 3,072 CPU only nodes
- 2 AMD "Milan" CPUs per node
- 512 GB memory per node
- Upgraded high speed network
- CPU partition will match or exceed performance of entire Cori system

Logos for NERSC, BERKELEY LAB, and U.S. DEPARTMENT OF ENERGY Office of Science are at the bottom.

NCAR: Cheyenne → Derecho



- NWSC-2: SGI ICE XA Cluster, 5.34PF
- CPU: Intel Xeon
- NO GPU!
- at the NCAR-Wyoming Supercomputing Center (NWSC) operational at the beginning of 2017.



- NWSC-3: HPE Cray EX cluster, 19.87PF
- CPU: AMD Milan (2488 Nodes)
- GPU: NVIDIA A100 (82 Nodes)
- **operational in Jan. 2023.**

- Cheyenneは“Conventional”なシステム。ユーザーの希望は、リスクより「枯れた」技術だった。
- 今回は調達前に利用者調査を行い、どこにどれだけ予算をかけるかを検討した。
 - 計算能力vsストレージ
 - HPC vs High-Throughput
 - CPU vs GPU
- 結果として、GPUが20%程度、という比率。

<https://www2.cisl.ucar.edu/sites/default/files/2021-10/Hart.pdf>

欧州のHPC

The European High Performance Computing Joint Undertaking (EuroHPC JU)

- 2018年設立, EUと加盟国33ヶ国による共同体
- “HPC : a key element for the EU’s digital strategy”

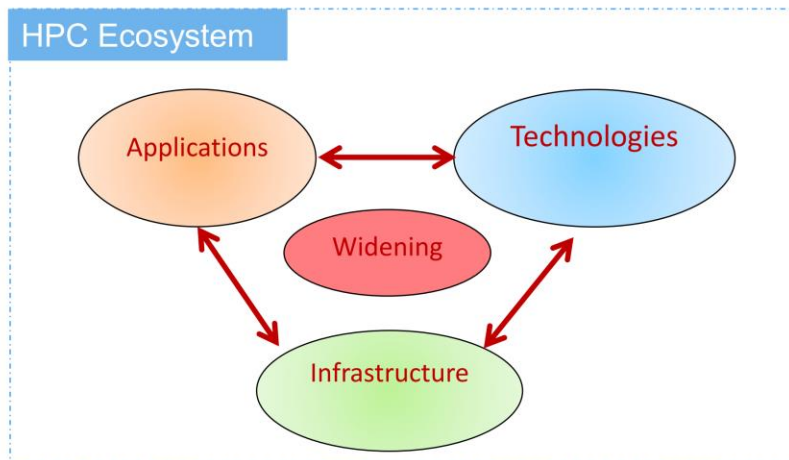
The EuroHPC Joint Undertaking 2019-2020



Mission: An integrated world-class supercomputing & data infrastructure and a highly competitive and innovative HPC and Big Data ecosystem

A legal and funding agency

- 33 Participating States (All EU MS) + EU + 2 Private Members (ETP4HPC & BDVA)
- Budget: 1.5 B€ (~1.1 B€ (536 m€ from EU) + In-kind from Private partners)



7

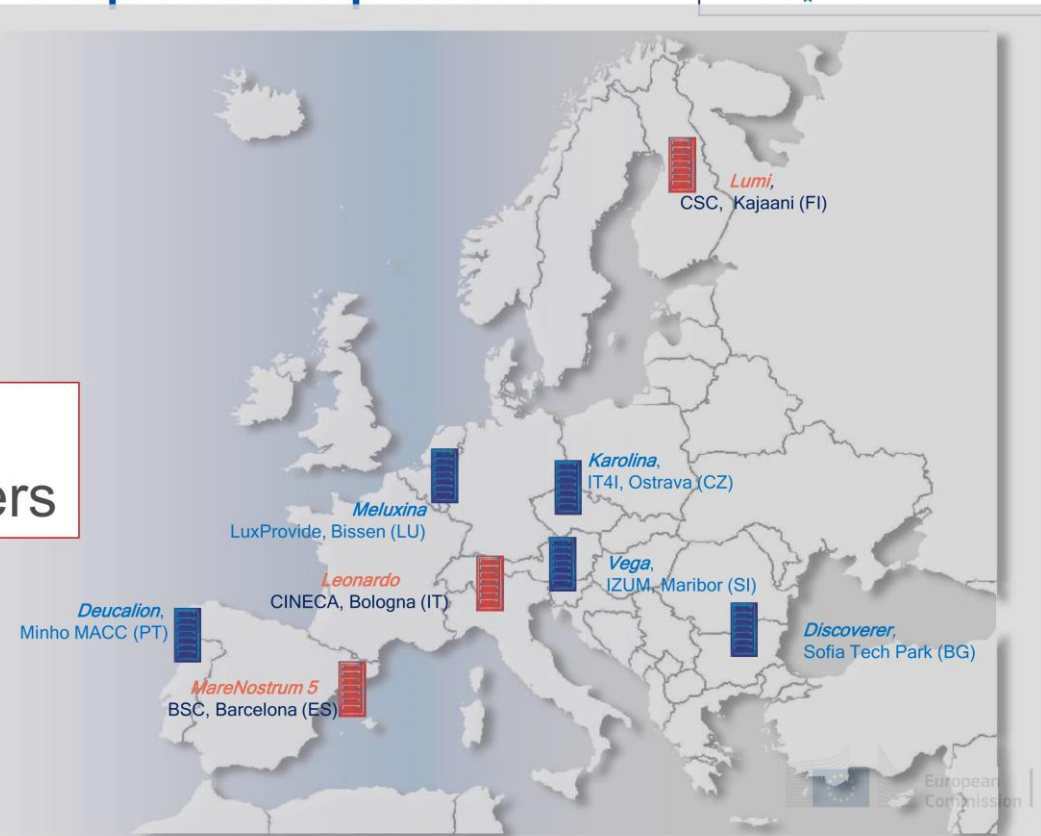
https://www.hpcuserforum.com/wp-content/uploads/2021/09/EuroHPC_Euro-Commission_L.Flores-Anover_Sept-2021-HPC-UF.pdf

Infrastructure – Supercomputers



Acquisition ~510 m€

Hosting entities of EuroHPC supercomputers



9

EuroHPC Pre-exascale system: 3つのPreExascale (+5つのPetascale)

- LUMI: 2022 full operation 予定
- 資源の半分はコンソーシアム10ヶ国で利用。
- 残りはEuroHPC JU経由で欧州研究者へ配分
 - PRACE のピアレビュープロセスと似た仕組み
- Cray EX、AMD EPYC CPU、AMD Instinct GPU、552 PF
- 100% hydropowered energy. Up to 200MWs are available.
- HIP、OpenMP5、Jupyter Notebook



LUMI, the Queen of the North

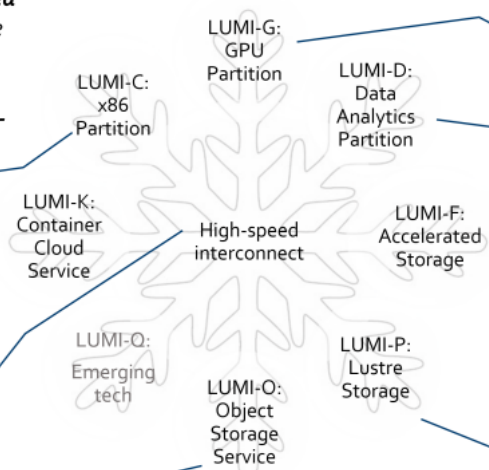
LUMI

LUMI is a Tier-0 GPU-accelerated supercomputer that enables the convergence of high-performance computing, artificial intelligence, and high-performance data analytics.

- Supplementary CPU partition
- ~200,000 AMD EPYC CPU cores

Possibility for combining different resources within a single run. HPE Slingshot technology.

30 PB encrypted object storage (Ceph) for storing, sharing and staging data



Tier-0 GPU partition: over 550 Pflop/s powered by AMD Instinct GPUs

Interactive partition with 32 TB of memory and graphics GPUs for data analytics and visualization

7 PB Flash-based storage layer with extreme I/O bandwidth of 2 TB/s and IOPS capability. Cray ClusterStor E1000.

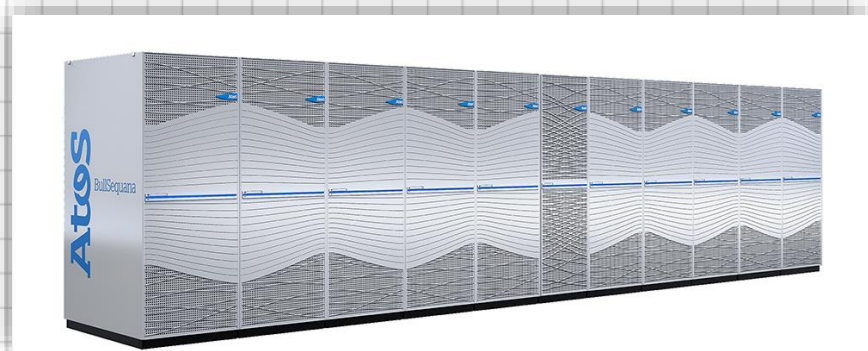
80 PB parallel file system

www.lumi-supercomputer.eu #lumisupercomputer #lumieurohpc

Leonardo@Italy, MareNostrum5@Spain

- Leonardo@CINECA (Italy)
- CINECA: 103の大学・公的機関からなるコンソーシアム
 - イタリアにおけるHPC、情報システムの開発・運用を担う
- will be installed at the end of 2022
- 323PF(Peak), ATOS BullSequana XH2000 CPU(Intel Ice lake/Sapphire), GPU(NVIDIA Ampere) Hybrid module
- 10 exaflops of FP16 AI performance.
- ECMWFの隣 (Tecnopolo di Bologna)

- MareNostrum5@BSC (Spain)
- スペイン、ポルトガル、トルコ、クロアチアがホスト
- 2021年6月に一旦入札中止、2022年1月再開。
- 2023運用開始予定
- 少なくとも205PF、GPUベース、CPUベースの2つのパーティション



- 気象・気候研究にどの程度使われるかは現段階では不明だが、CINECAはイタリア国内の気象・気候研究の中心でもあるので、ある程度は利用されるのではないか。
- BSCはIS-ENES3でも中心的役割を果たしているなので、間違いなく活用されると思われる。

その他欧州の気象気候の現業・研究機関

ECMWF

- 2021.09 New data center, Bull Sequana XH2000-based
- four “self-sufficient … clusters,”
- AMD EPYC (Rome), NO GPU.
- repurposed tobacco factory



DKRZ

- HLRE-3 “Mistral”, 2015–2022
 - Bull DLC720, 3.6PF, Haswell/Broadwell, NO GPU.
- HLRE-4 “Levante”, 2021–
 - Atos BullSequana XH2000, 16+PF
 - CPU: AMD EPYC Millan, GPU: Nvidia A100
- ICON-A以外にGPU対応しているコードがない!
 - → 1ラックだけGPUノードを導入



UKMO

- “Supercomputer as a Service” with Microsoft Azure on HPE Cray EX
- AMD EPYC CPU
- 1.5 million processor cores and over 60 petaflops,
- becoming operational starting July 2022.

DWD

- NEC SX-Aurora 5.6 / 4.3PF (Research用 / 現業用), 2019–2023
- 現業モデルがCOSMOからICON-LAMに代わる。
 - → 「次はGPUかもしれない。」

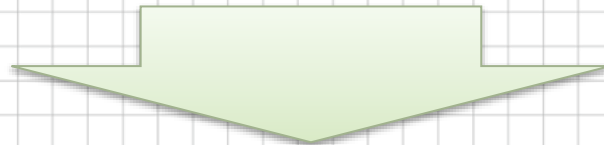
現業センターはトップエンドのHPCを追いかけるだけという訳にはいかない。

まとめ

まとめ

- 米国のトップエンドHPCはExascale目前
- 欧州はやっとPre-Exa。ただし、3つ。
- いずれも、CPU-GPUハイブリッドが主。ただし、CPU (Intel、AMD)・GPU (NVIDIA、AMD、Intel)の組合せが多様化。
- 現業機関はトップエンドからは一歩二歩後をついていく様子。

これからもこれまで同様、HPCは気象気候分野に不可欠。
だが、今後はHPC業界は気象気候分野に特別に注力する必要はない。



汎用に作られた道具、最先端の道具を自分たちの研究のために使いこなそうとおもえば、そのための能力と努力が必要。



Thank You.