슈퍼컴퓨팅과 고성능컴퓨팅(HPC)

넥스트폼

이 보 성
bs.lee@nextfoam.co.kr

슈퍼컴퓨팅(Supercomputing)의 정의

- 슈퍼컴퓨팅(Supercomputing)
  - 병렬로 작동하는 다중 컴퓨터 시스템(예: "슈퍼컴퓨터")의 집중된 컴퓨팅 리소스를 사용하여 매우 복잡한 문제나 데이터 집약적인 문제를 처리하는 것

- 슈퍼컴퓨터(Supercomputer)
  - 슈퍼컴퓨터는 과학기술연산을 비롯한 다양한 분야에 사용되는 고속/거대 용량 컴퓨터
  - 일반적 목적의 컴퓨터에 비해 당대 최상급 처리 능력 (특히 연산 속도)을 보유한 고성능 컴퓨터로, 절대적 기준이 아닌 상대적인 개념

- 고성능컴퓨팅(High Performance Computing, HPC)
  - 복잡한 연산 문제를 풀기 위하여 슈퍼컴퓨터 및 컴퓨터 클러스터를 사용하는 것 (위키백과)

전 세계의 대표적인 슈퍼컴퓨터

- Top500 (https://top500.org/) 순위
  - 매년 6월, 11월에 발표하는 전세계 슈퍼컴퓨터의 성능 순위
  - HPCG(High-Performance Conjugate Gradient)와 에너지 효율성 순위인 Green500 순위도 발표

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016

2023. 6월 Top500 1 ~ 3 위 슈퍼컴퓨터 (https://top500.org/lists/top500/2023/06/)

https://www.ornl.gov/news/ornl-celebrates-launch-frontier-worlds-fastest-supercomputer

한국 대표 슈퍼컴퓨터

- 국가슈퍼컴퓨팅센터 (한국과학기술정보연구원, http://ksc.re.kr)

1호기: Cray-2S, 1988
CPU 4개, 1GB 메모리

2호기: Cray C90, 1993
CPU 16개, 4GB 메모리

3호기: IBM p690 (CPU 672개)
NEC SX-5, 6, 2001~2002

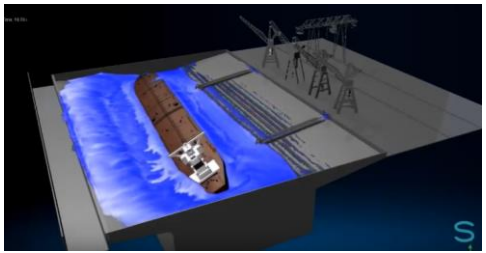
4호기: Sun Linux Cluster (AMD),
2010

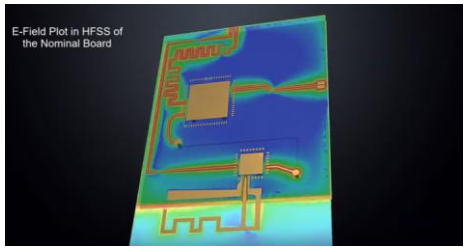
5호기: Cray Linux Cluster (Intel), 2018
2019년 11월 기준 세계 14위

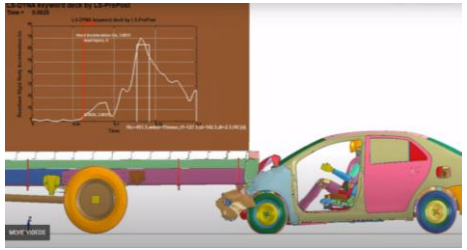
KISTI

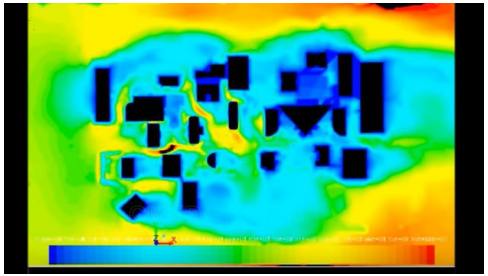
슈퍼컴퓨팅(고성능 컴퓨팅) 활용 분야

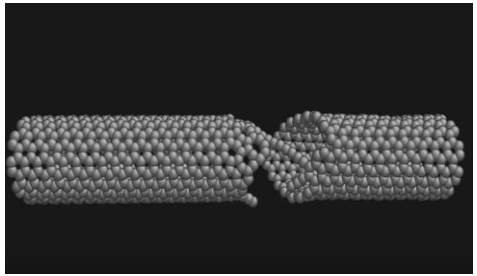
- 장시간, 고성능, 대용량의 컴퓨팅 자원이 필요한 시뮬레이션 (Big Problem)
  - 일기 예보, 기상 연구, 단백질 입체 구조 예측, 양자 역학, 분자 동역학, 소재 개발
  - 항공기, 자동차의 공력 및 충돌 해석, 핵폭발 및 핵융합의 연구, 우주 연구, 풍공학
- 확률 및 통계 연산, 경기 예측, AI, ChatGPT

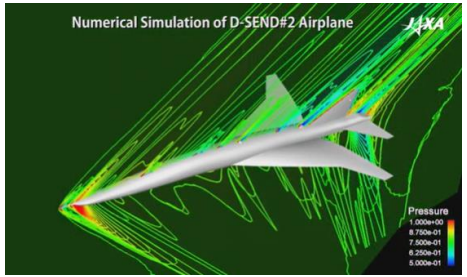












클라우드 컴퓨팅과 고성능 컴퓨팅

- 클라우드 컴퓨팅의 정의
  - 하드웨어, 소프트웨어, 데이터 등 IT 자원을 네트워크를 통해 표준화된 서비스(IaaS, SaaS, PaaS) 형태로 제공하는 컴퓨팅 모델

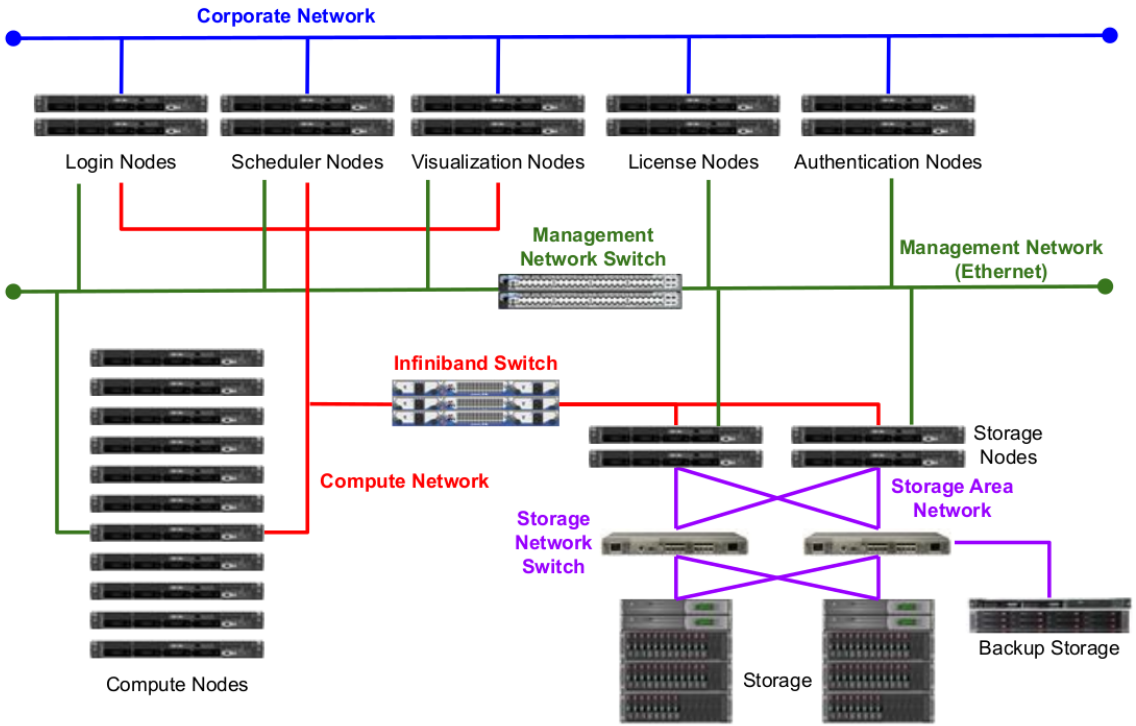
- HPC 클라우드
  - HPC 자원을 클라우드 컴퓨팅에서 제공 (HPCaaS)
  - 2023.6 Top500의 11위에 MS Azure HPC가 등재

11	**Explorer-WUS3** - ND96_amsr_MI200_v4, AMD EPYC 7V12 48C 2.45GHz, AMD Instinct MI250X, Infiniband HDR, Microsoft Azure West US3 United States	445,440	53.96	86.99

HPC 슈퍼컴퓨터 구성 요소

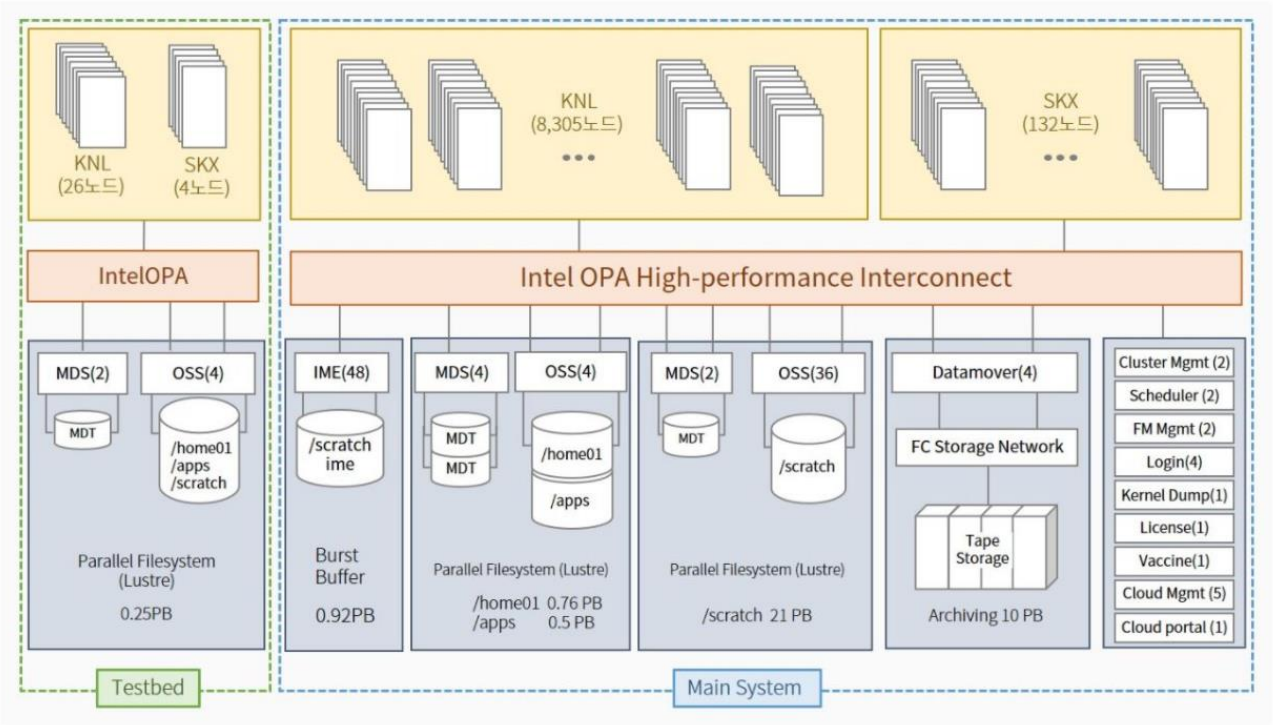
- Infrastructure stack
  - Processor / Cache / Memory
    - CPU Intensive / Memory Intensive / GPU
  - Interconnect network
    - Ethernet / RDMA (Infiniband or RoCE) / EFA(Elastic Fabric Adapter)
  - Storage
    - Local block storage / NFS / Parallel file system / Object storage
- Management stack
  - Job management / Resource management / Monitoring / Security
- Applications stack
  - Commercial applications
  - Open source applications
  - Custom / In-house applications & containers

일반적인 HPC 클러스터 구성



1. User / Management Servers
   - Login nodes
     - User login / Job submission
   - Scheduler nodes
     - LSF / PBS / Slurm
   - Visualization nodes
     - NiceDCV / VNC / RD
   - License nodes
   - Authentication nodes
     - LDAP / AD / NIS
2. Compute Nodes
3. Management Network
   - Connects all nodes in the cluster
   - Management / monitoring
4. Compute Network
   - MPI communication & File I/O
5. Mountable Storage System
   - Storage Nodes
     - NFS / Lustre / BeeGFS
   - Storage Area Network
     - Storage system internal network
   - Storage Disk
     - HDD / SSD / FC Disk
6. Backup Storage

슈퍼컴퓨터 급 HPC 클러스터 – KISTI Nurion



1.Intel OPA (Omni-Path Architecture)
- Compute network
- File I/O network

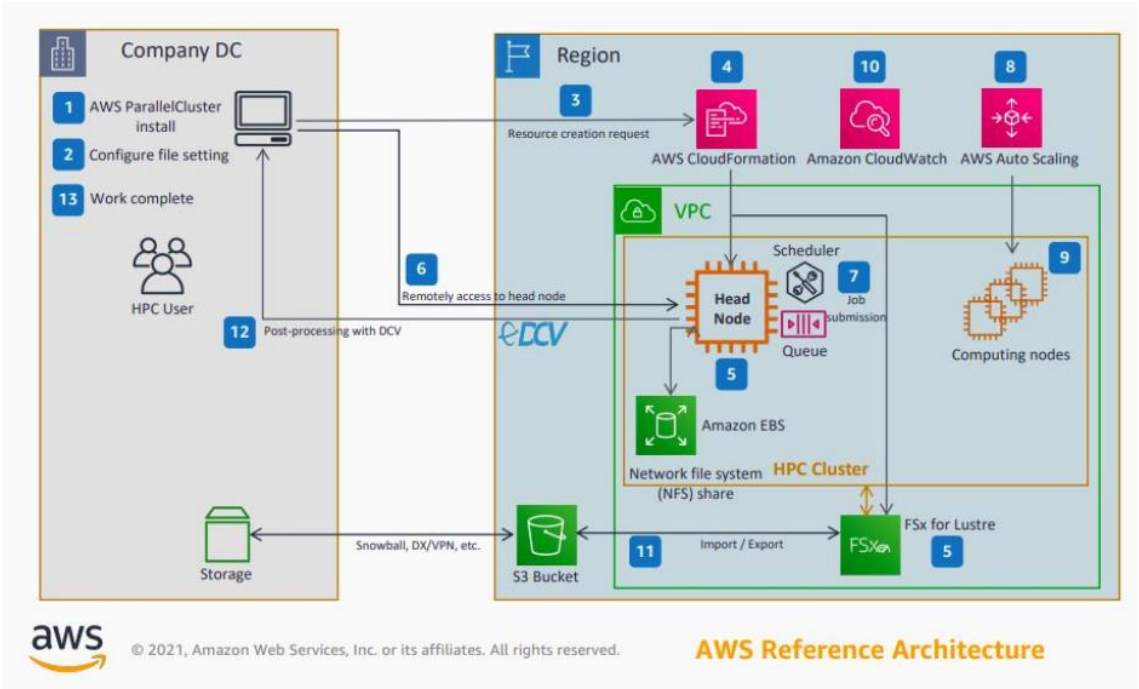
2. Parallel File System
- /home01, /apps
- /scratch

3. Burst Buffer Storage
- NVMe based IO cache between compute nodes and parallel file system
- improve the performance of small or random I/O in the Lustre
- DDN IME240

4. Testbed system
- Code test, development
- Interactive run

Lift-and-Shift HPC Cloud

- On-Premise HPC를 클라우드 환경으로 그대로 들어서 옮긴 HPC 클라우드



HPC 클러스터의 성능 요인 – Processor 와 캐시메모리

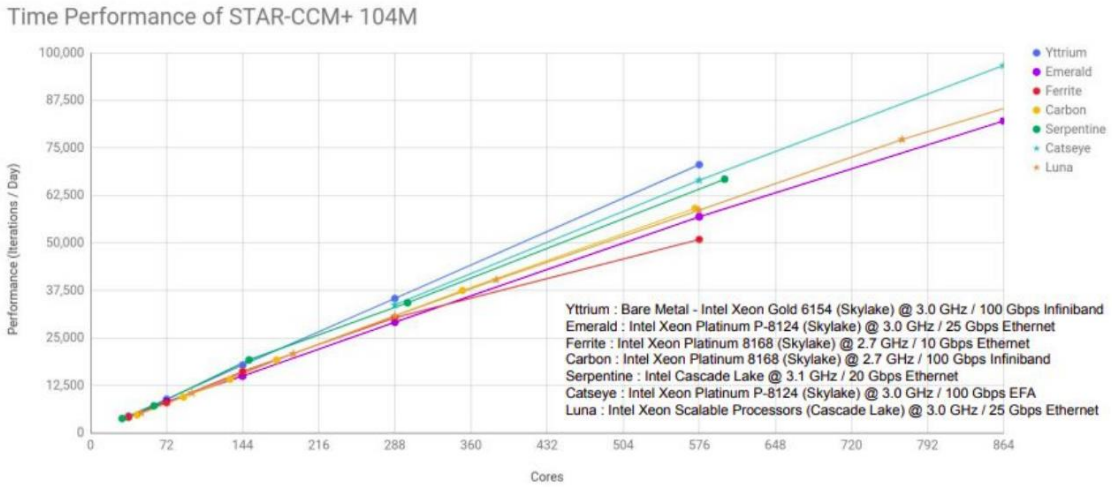
- CPU의 성능 뿐만 아니라 CPU와 메모리 간의 Bandwidth, 캐쉬 메모리 등이 성능에 영향

Processor	Base Clock (GHz)	캐시 메모리 (MB)	Cores / Node	Analysis	Duration (s)	GFlops
AMD EPYC 7742 (Rome)	2.45	256	60 / 128	StarCCM+ 17.02.008	345	1564.9
AMD EPYC 7V3X (Milan-X)	2.2	768	64 / 128	StarCCM+ 17.02.009	246	1951

Processor	캐시 메모리 (MB)	STREAM (Gb/s)	Memory Copy (Gb/s)	Dumb MBW (Gb/s)	Memory Block (Gb/s)	B_eff (Gb/s)
AMD EPYC 7742 (Rome)	256	272	15.2	14.7	9.2	59.2
AMD EPYC 7V3X (Milan-X)	768	278.1	21.3	21	14.9	65

HPC 클러스터의 성능 요인 – 인터커넥트 네트워크

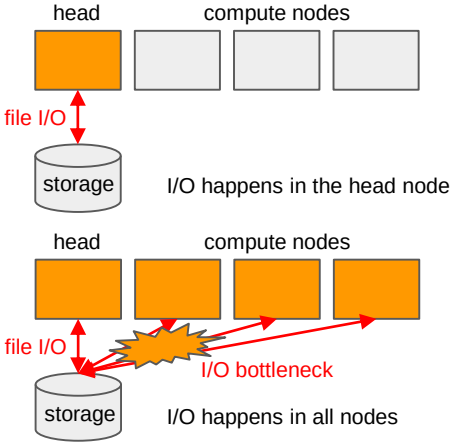
- 인터커넥트 네트워크
  - HPC 클러스터의 컴퓨팅 노드간 네트워크로 Bandwidth보다 Latency가 더 큰 영향을 끼침



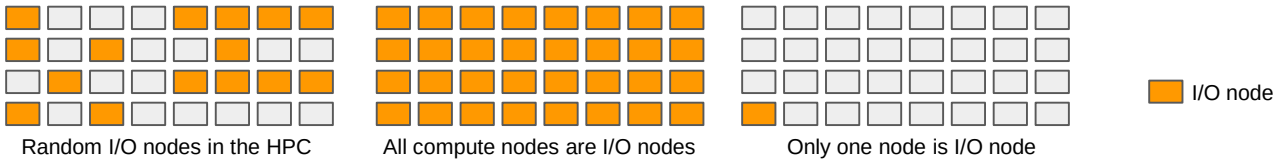
- Catseye : Emerald - Same processor / 100 Gbps EFA vs 25 Gbps non-EFA
- Carbon : Ferrite - Same processor / 10 Gbps Ethernet vs 100 Gbps Infiniband
- Yttrium : Catseye - Similar processor / Bearmetal vs Virtuallized

HPC 클러스터의 성능 요인 – 스토리지

- 시뮬레이션 워크로드에 따라 I/O 패턴이 다양
  - I/O occurs in the head node only
    - Almost CFD/FEA simulations
  - Parallel I/O in all compute nodes
    - WRF, MPI-IO code, Fluent Parallel I/O etc
    - EDA application exchange data using the storage

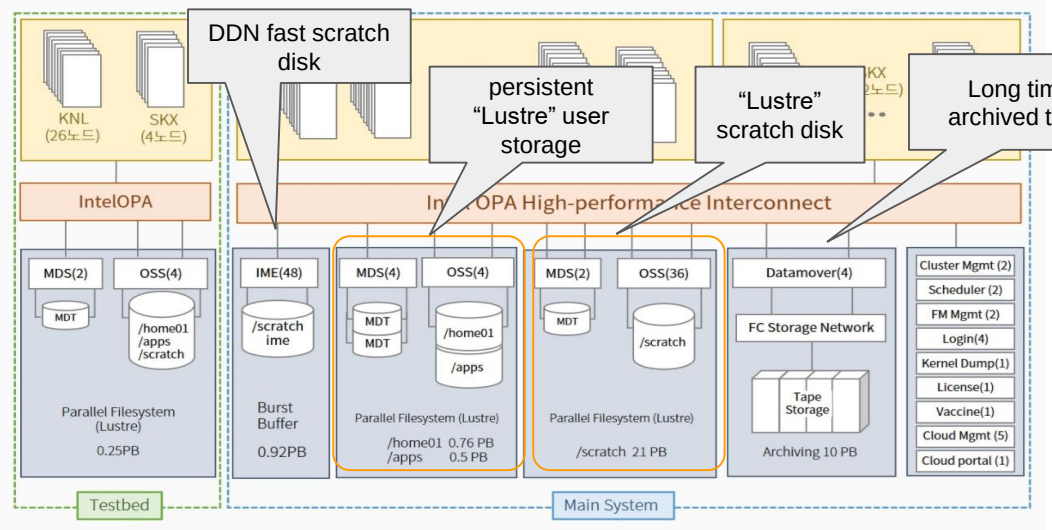


- 다양한 스토리지 요구사항이 존재
  - All compute nodes can access the shared storage without I/O bottleneck for the scratch
  - All user data should be stored in the persistent shared storage



슈퍼컴을 위한 스토리지 구성

- 다양한 요구사항을 충족하기 위한 스토리지 구성 필수
  - High performance parallel I/O for the fast scratch disk
  - Persistent shared large sized storage for user data
  - Long time archived storage



KISTI Nurion conceptual architecture

PFSs are used differently in a single HPC

고성능 슈퍼컴퓨터 설계 방안

- General purpose HPC cluster
  - Dedicated homogeneous instances
  - RDMA enabled interconnect for MPI applications
  - Parallel file system with low latency storage network for the scratch and user directory
  - Job scheduler and module based application management
  - Auto-scaling for the bursting situation

- Design for the major applications
  - CPU Intensive / memory intensive / GPU applications
  - Interconnect network is dependent on the application characteristics
  - For non-MPI applications but distributed computing, standard Ethernet maybe sufficient
  - Non parallel file system but shared storage system for user directory
  - Job scheduler may not be needed – FIFO