Hewlett Packard Enterprise

HETEROGENOUS ARCHITECTURE FOR MIXED HPC AND AI WORKLOAD FOR PUBLIC RESEARCHER AND PRIVATE SECTORS

Dr Bruno Leconte
Southern Europe & CEEMA HPC & AI Solution Architect Group Manager
Hewlett Packard Enterprise

September 19, 2019

INFORMATION & ACKNOWLEDGEMENTS

- **HPE makes no warranties regarding the accuracy of this information**. This document may contain forward looking statements regarding future operations, product development, product capabilities and availability dates. This information is subject to substantial uncertainties and is subject to change at any time without prior notification. Statements contained in this document concerning these matters only reflect Hewlett-Packard Enterprise's predictions and / or expectations as of the date of this document and actual results and future plans of Hewlett-Packard Enterprise may differ significantly as a result of, among other things, changes in product strategy resulting from technological, internal corporate, market and other changes. **The information contained in this presentation** is proprietary to Hewlett Packard Enterprise (HPE) Company or to it's respective owners and is offered in confidence, subject to the terms and conditions of use for RES 2019 event.

- Acknowledgements:
  - Denis Girou, Director of IDRIS
  - Bruno Monnet, HPC & AI Senior consultant @ HPE
  - Pierre Eric Bernand, HPC & AI Senior Application Engineer @ HPE
  - Jin Shin, AI Senior Application Engineer @ HPE

FRANCE'S AI STRATEGY

- The French government plans to invest €1.5 billion over five years to support research and innovation in the field of AI
- Country's strategy in the field of artificial intelligence (AI) Key proposal :
  - Developing an aggressive data policy
  - Targeting four strategic sectors
  - Boosting the potential of French research
  - Planning for the impact of AI on labour
  - Making AI more environmentally friendly
  - Opening up the black boxes of AI
  - Ensuring that AI supports inclusivity and diversity

https://www.aiforhumanity.fr/

AI FOR HUMANITY

FRENCH STRATEGY FOR ARTIFICIAL INTELLIGENCE

The President of the French Republic presented **his vision and strategy** to make France a **leader in artificial intelligence (AI)** at the Collège de France on **29 March 2018.**

Download the Villani Report

© Elysée

FRANCE'S AI STRATEGY

- Villani's report (page 75 - https://www.aiforhumanity.fr/en/) :

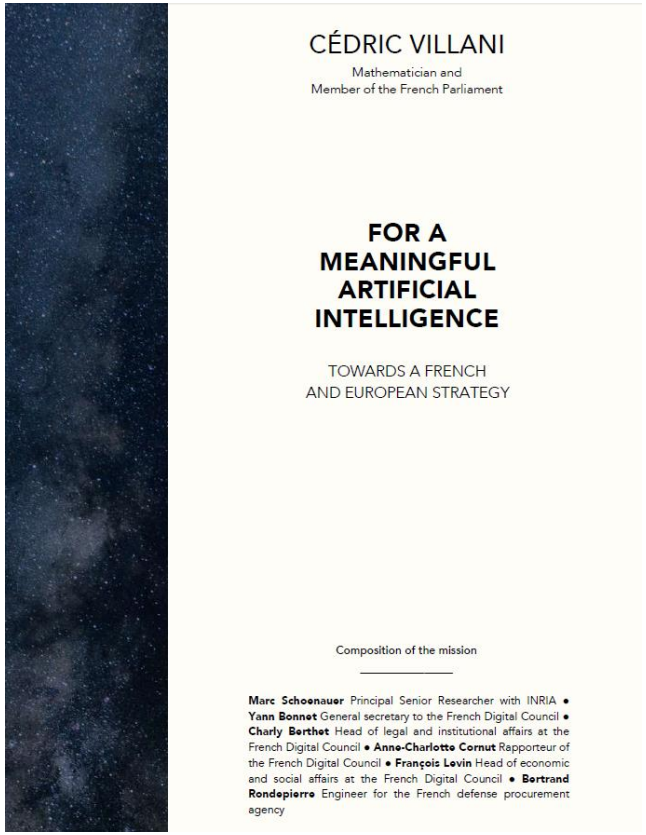
"Several types of work flow in the learning stage do need distinguishing, however: cases which will for the **most part rely on a supercomputer adapted to AI** (resources typically entailing thousands of GPUs) are fairly rare and only concern part of the research focusing on AI. The vast majority of applications require a much smaller equipment setup (entailing a few dozen GPUs for example). These two complementary requirements should not be confused but kept distinct, therefore, as they are very different in nature and setup:

- A **requirement in terms of supercomputer** that is designed and dedicated entirely to AI;
- **A requirement in terms of "cloud adapted to AI"**, the beneficiaries of which will include research.

...

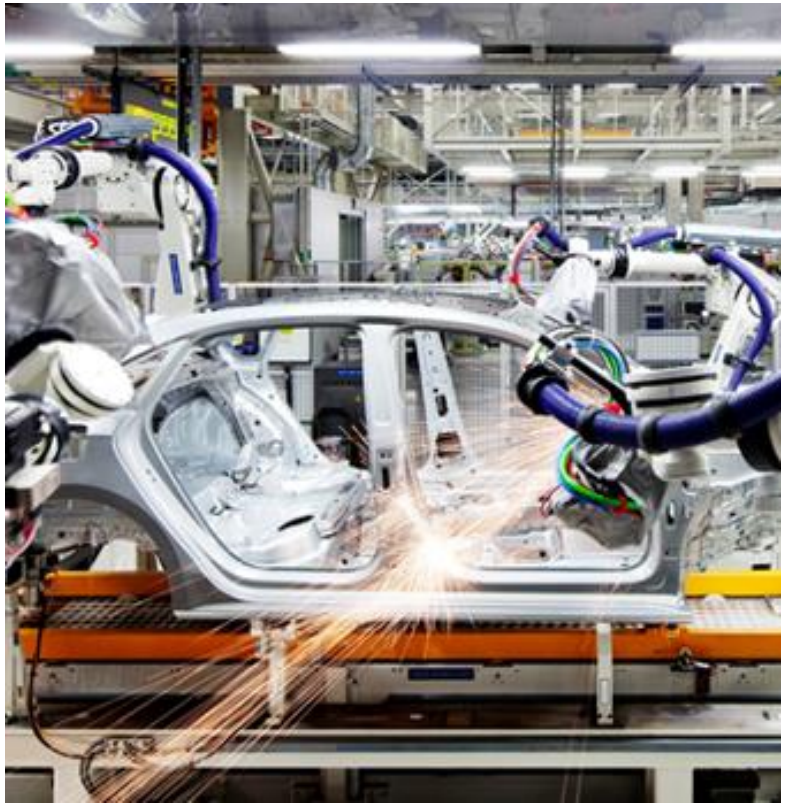
- Developing a Supercomputer for the Requirements of Research
  - The recommendation is to **set up a supercomputer designed specifically for artificial intelligence applications** (such architecture differs significantly from conventional HPC supercomputers), solely for use by French research, beginning with the members of the 3IA institutes, described above, and their industrial partners under joint projects. By limiting access to free access for research is the only way to achieve an access mechanism that is simple in both administrative terms and in terms of everyday use. For opening up broader access on a paying basis for some users ends up attaching conditions to use of the tool in practice.
  - **Private businesses specializing in the field** will have to be **called on to design such infrastructure** and specifications will have to be drawn up that are specific to AI.

"



HPE VISION FOR AI & HIGH PERFORMANCE COMPUTING

Advancing the way people live and work with HPC and AI Solutions



$1 B investment on HPC & AI technologies

Flexible Consumption Models

ACCELERATING OUR STRATEGY THROUGH PARTNERSHIPS

Innovations

Composable Infrastructure
HPC & AI compute
Secure compute experience

All-Flash Storage
In-memory analytics
Memory-Driven Computing

Mobile-First Wireless
Edge Computing
Location-based Services

Recent acquisitions

CLOUD CRUISER
Cloud Technology Partners
PLEXXI
RedPixie
CRAY
sgi
simplivity
bluedata
nimblestorage
aruba a Hewlett Packard Enterprise company
cape
niara
RASA

Expanded Partnerships

Hybrid Cloud

accenture
ARISTA
docker
DXC.technology
HORTONWORKS
MESOSPHERE
MICRO FOCUS
Microsoft
NVIDIA
SAP
tamr
VEEAM
vmware
wipro

AI & Advanced Analytics

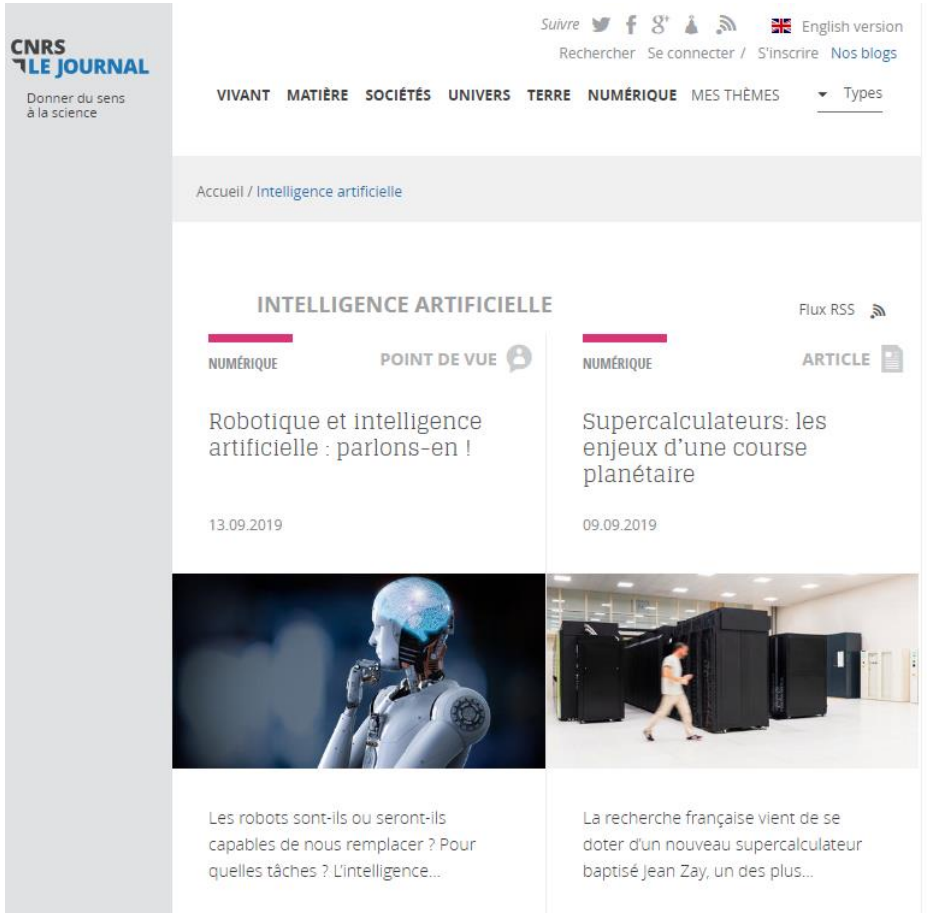
Caffe
Caffe2
cassandra
data iku
druid
kafka A distributed streaming platform
kinetica
nifi
NVIDIA
PyTorch
Qlik
TensorFlow

Redefining experiences at the Edge

AT&T
CITRIX
Deloitte.
GE Digital
hp
Microsoft
NATIONAL INSTRUMENTS
paloalto NETWORKS
pwc
ptc
Tech Mahindra

FRANCE'S SUPERCOMPUTER DEDICATED TO MIXED HPC AND AI WORKLOAD

- Jean Zay* is the new supercomputer converged platform acquired by the French Ministry of Higher Education, Research and Innovation through the intermediary of GENCI (Grand Equipement National de Calcul Intensif)
- Schedule:
  - **Acquisition** was signed into contract with Hewlett Packard Entreprise (HPE) on 8 **January 2019**
  - Installation 1st half of CY 2019
  - **Big Challenges projects during summer 2019**
  - Opening to all users at the end of summer 2019
- Location:
  - **IDRIS**, National Computing Centre for the French National Centre for Scientific Research (CNRS**).



© C. Frésillon/ IDRIS / CNRS Photothèque

* In honour of Jean Zay (1904-1944), French Education and Arts Minister from June 4, 1936 to September 10, 1939

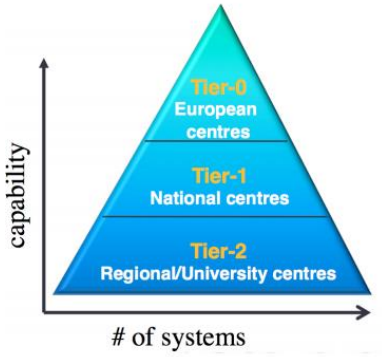
** CNRS : Founded in October 1939, 32 000 employees, 11 000 researchers, budget 3,3 B€

FRANCE'S SUPERCOMPUTER DEDICATED TO MIXED HPC AND AI WORKLOAD

- IDRIS :
  - Institut du Développement et des Ressources en Informatique Scientifique
  - Operating since 1993
  - IDRIS is one of the three **Tier-1 French sites**
  - IDRIS is the main HPC center of CNRS
    - a computing resource center with state of the art high end hardware and software
    - a center of excellence in information technologies applied to computational science
  - With Jean Zay supercomputer, IDRIS will **continue to serve the HPC community** but will become additionally the **first national platform for AI usage**
  - Located on the Paris-Orsay university campus (South of Paris)

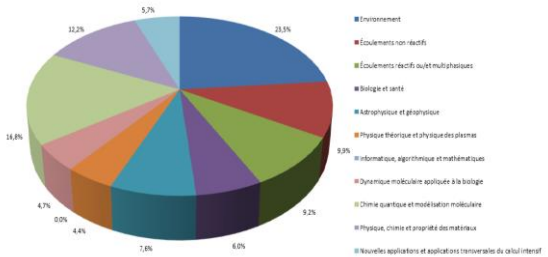
© C. Frésillon/ IDRIS / CNRS Photothèque

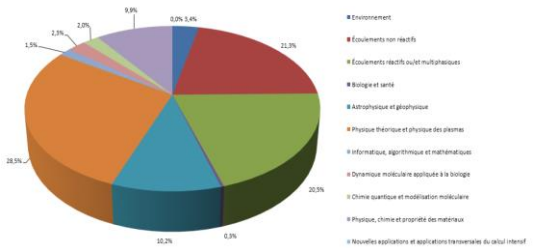
Tier-0: European Centres (> petaflop machines)
Tier-1: National Centres
Tier-2: Regional/University Centres



FRANCE'S SUPERCOMPUTER DEDICATED TO MIXED HPC AND AI WORKLOAD

- IDRIS Users :
  - Public research agencies
  - Universities
  - Industries for open research projects
- Users access steps :
  - Evaluation Committee
  - Attribution Committee
  - Up to now, IDRIS manages around 250 HPC scientific projects and 1000 users





- 11 dedicated scientific committees (with a total of around 120 experts) :
  - 1. Climate, Environment
  - 2a. Not reactive flows
  - 2b. Reactive or multiphasic flows
  - 3. Biology and Health
  - 4. Astrophysics and Earth Sciences
  - 5. Theoretical Physics, Plasmas
  - 6. Computer Sciences, Mathematics
  - 7. Molecular dynamics for Biology
  - 8. Quantum Chemistry et Molecular Simulation
  - 9. Material Physics
  - 10. Multi-thematic Applications, Emergent Applications, Artificial Intelligence

FRANCE'S SUPERCOMPUTER DEDICATED TO MIXED HPC AND AI WORKLOAD

- IDRIS HPE SGI 8600 HPC & AI supercomputer:
  - Scalar component :
    - Peak performance 5.08 PFlop/s
    - 1528 XA730i nodes, 61 120 Intel 6248 cores
    - For each node
      - 2 Intel Cascade Lake 6248 processors with 20 cores at 2.5 GHz
      - 192 GB of DDR4 memory 2666 MT/s (4.8 GB/core)
      - 1 Intel OmniPath (OPA) 100 Gb/s link
  - Hybrid converged component :
    - Peak performance 9.05 PFlop/s
    - 261 hybrid nodes XA780i, 10 440 Intel 6248 cores,
    - 4 NVIDIA V100 32 GB GPUs per node, total 1 044 GPUs
    - 4 Intel OmniPath (OPA) 100 Gb/s link by node
  - Full flash parallel filesystem : 1 PB & 300 GB/s with 1.16 million IOPS
  - Intel OPA 100 Gb/s network in *Enhanced Hypercube* topology

© C. Frésillon/ IDRIS / CNRS Photothèque

HETEROGENOUS ARCHITECTURE FOR MIXED HPC AND AI WORKLOAD

Get IDRIS & GENCI AI ressources to users

- Communicate : let users know it is available
- Understand user's need :
  - survey
  - new research area (sociology, computer vision, ML, natural langage processing, digital entertainment, Music, virtual human…)
- Setup access to the new AI supercomputer that meet users requirements :
  - standard HPC access
  - jupyter notebook,
  - containers…

DEEP LEARNING FRAMEWORKS AND THEIR DEPENDENCIES

UI, development tools	jupyter NVIDIA DIGITS (Caffe, Torch, TensorFlow)
High-level APIs	Keras (TensorFlow, CNTK, Theano, MXNet) TF Layers (TensorFlow) Brew (Caffe2) GLUON (MXNet) RAPIDS
Deep learning and machine learning frameworks	TensorFlow Caffe CNTK NCCL data iku PyTorch theano scikit learn Caffe2 BigDL mxnet KALDI H2O.ai
Hardware-specific libraries for basic operations for deep neural networks (BLAS + FFT, convolutions, etc)	cuDNN Intel® MKL-DNN MIOpen
Optimized linear algebra libraries, many support BLAS interface, hardware specific	cuBLAS, MKL, OpenBLAS, rocBLAS, MIOpenGEMM
Accelerator-specific drivers and software	NVIDIA drivers, CUDA, ROCm

HETEROGENOUS ARCHITECTURE FOR MIXED HPC AND AI WORKLOAD

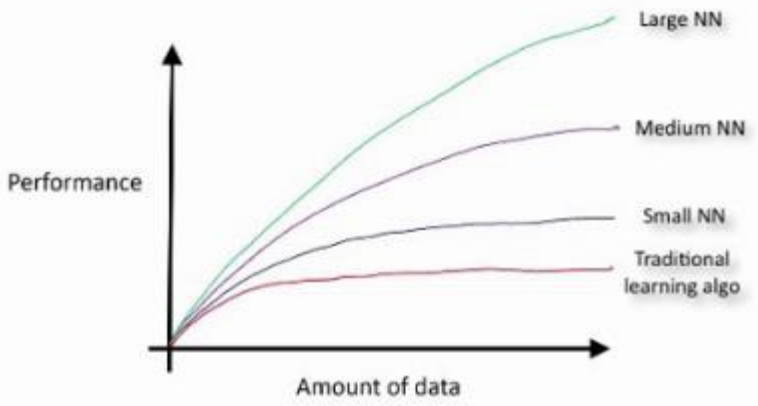
- What an HPC center can offer to users with heterogenous architecture for mixed HPC and AI workload “today” ?

HPE SGI 8600

Large supercomputer for HPC and AI

Chainer, Keras, Caffe, Caffe2, HOROVOD, PYTORCH, mxnet, Microsoft CNTK, TensorFlow, KALDI

Frameworks & high level API



Andrew Ng

HETEROGENOUS ARCHITECTURE FOR MIXED HPC AND AI WORKLOAD

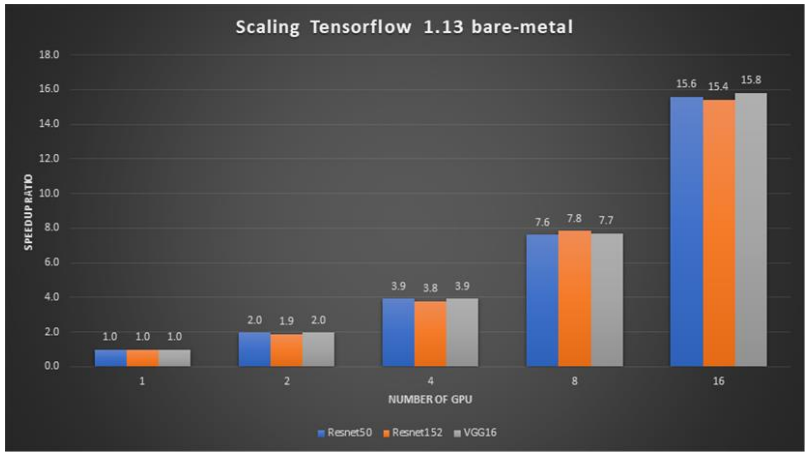
- What an HPC center can offer to users with heterogenous architecture for **performance** ?

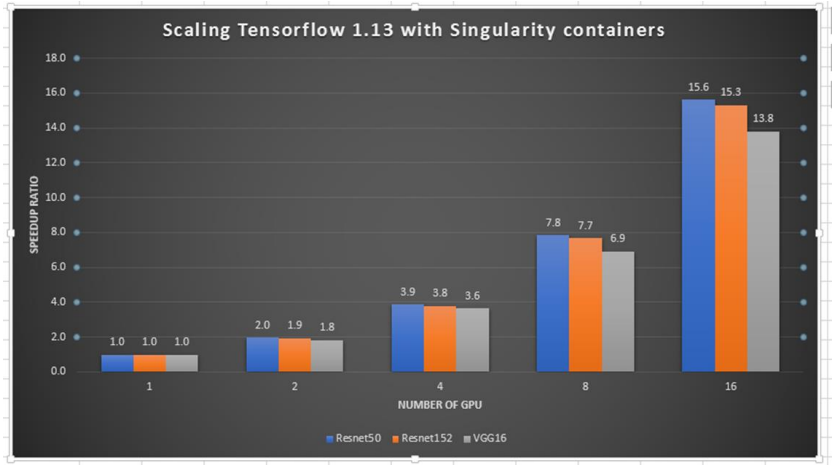
HPE SGI 8600

Large supercomputer for HPC and AI

Chainer, Keras, Caffe, Caffe2, HOROVOD, PYTORCH, mxnet, Microsoft CNTK, TensorFlow, KALDI

Frameworks & high level API





HETEROGENOUS ARCHITECTURE FOR MIXED HPC AND AI WORKLOAD

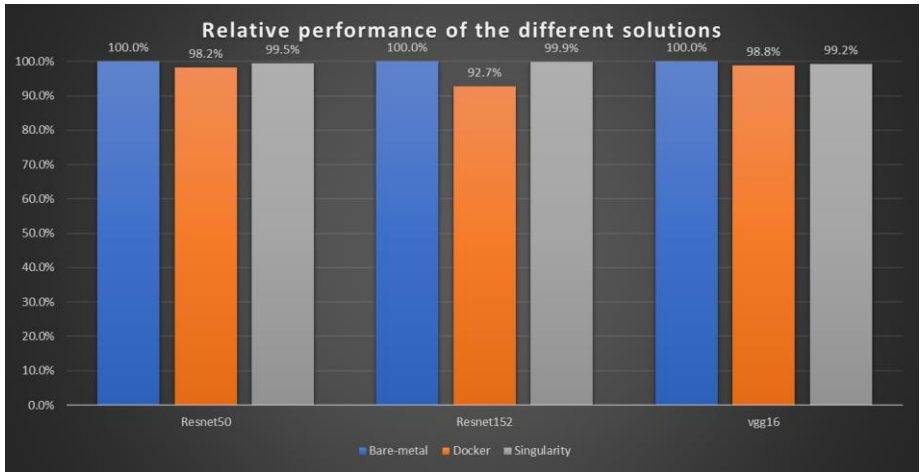
- What an HPC center can offer to users with heterogenous architecture for **performance** ?

HPE SGI 8600

Large supercomputer for HPC and AI

Chainer Keras Caffe Caffe2 HOROVOD PYTORCH mxnet Microsoft CNTK TensorFlow KALDI

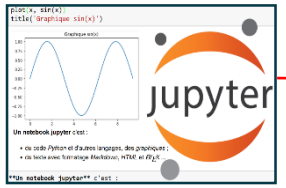
Frameworks & high level API



HETEROGENOUS ARCHITECTURE FOR MIXED HPC AND AI WORKLOAD

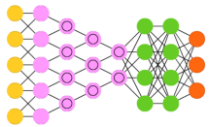
- Large HPC center *Work In Progress* for HPC and AI workload ?

HTTPS



HTTPS

HPE SGI 8600



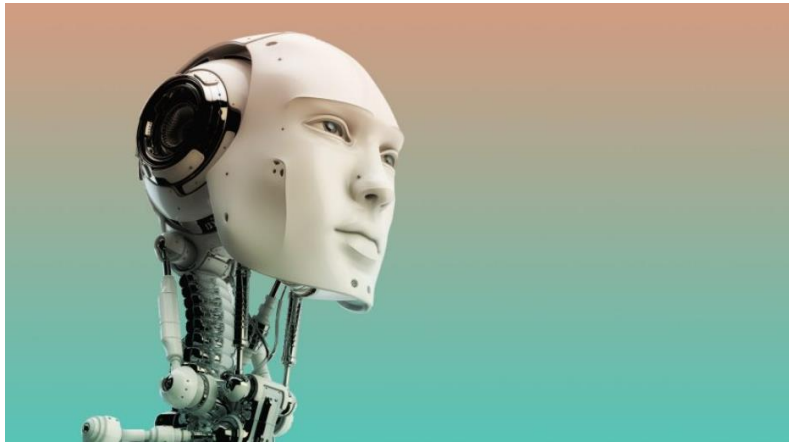
WIP

Thousands of containers
+
http reverse proxy

Large supercomputer for
HPC and AI

Q&A

Hewlett Packard Enterprise



Jean Zay HPE SGI 8600 - http://www.idris.fr/annonces/annonce-jean-zay.html

Heterogenous architecture for mixed HPC and AI workload for public researcher and private sectors