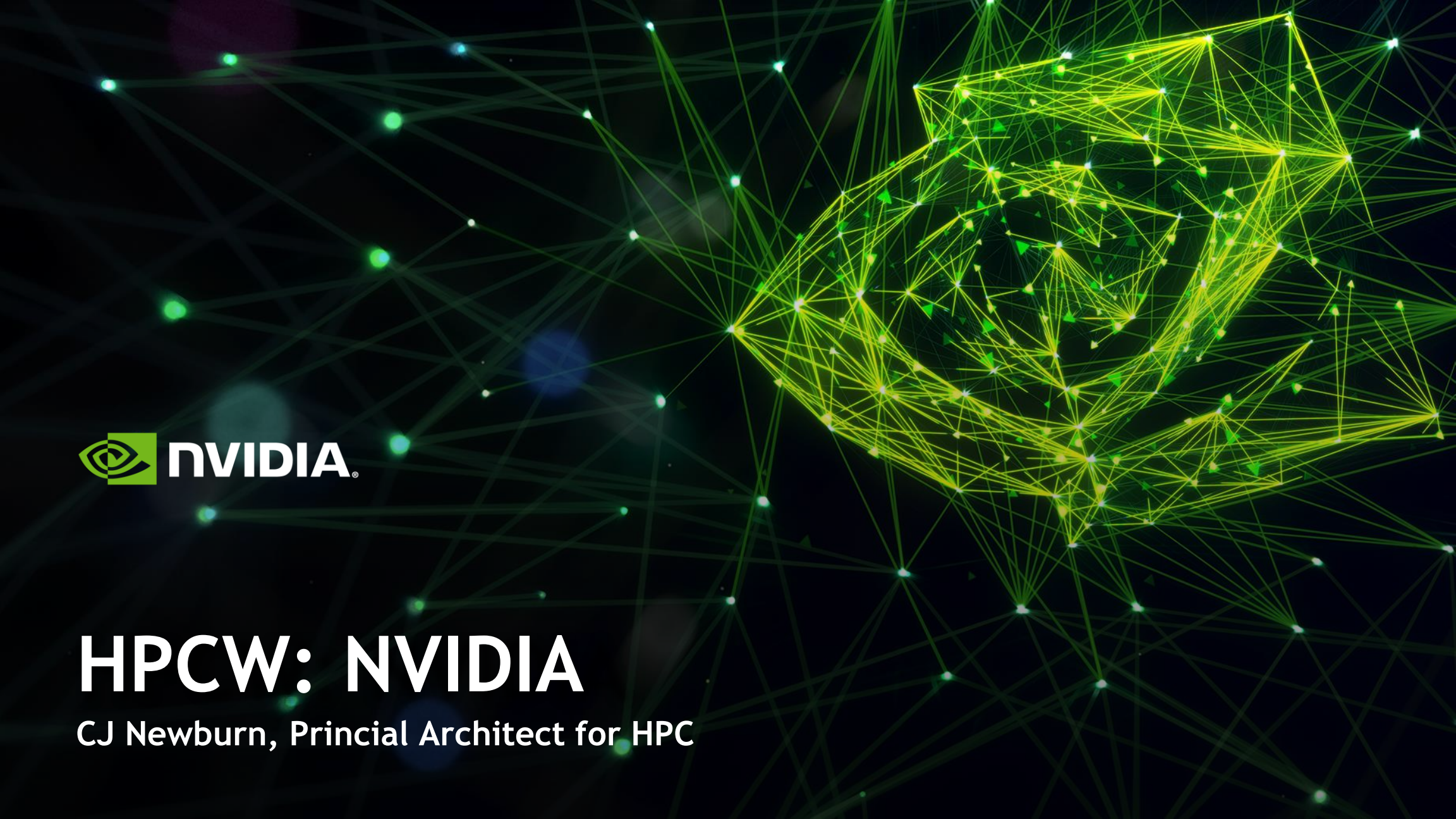


HPCW: NVIDIA

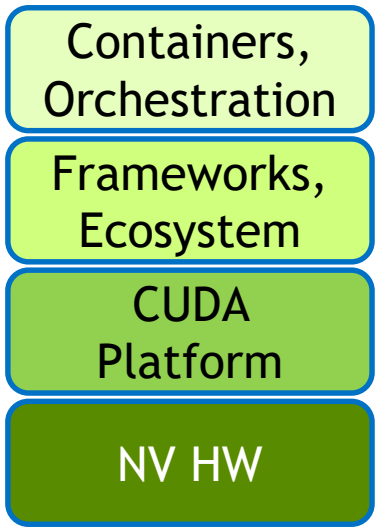
CJ Newburn, Princial Architect for HPC

NVIDIA OVERVIEW

- Containers and HPC
- What NVIDIA is doing
- NVIDIA GPU Cloud

WHY CONTAINER TECHNOLOGIES MATTER TO HPC

Good for the community, good for NVIDIA



- Democratize HPC
  - Easier to develop, deploy (admin), and use
- Good for the community, good for NVIDIA
  - **Scale** → HPC; more people enjoy benefits of our scaled systems
  - Easier to deploy → less scary, less complicated → **more GPUs**
  - Easier to get all of the right ingredients → **more performance** from GPUs
  - Easier **composition** → HPC spills into adjacencies

WHAT NVIDIA IS DOING

Earning a return on our investment

- Container images, scripts, and industry-specific pipelines in **NGC** registry
  - Working with developers to tune scaled performance
  - Validating containers on NGC and posting them in registry
  - Used by an increasing number of data centers
- Making creation and optimization automated and robust with **HPCCM** (blog)
  - Used for every new HPC container in NGC, broad external adoption
  - Apply best practices with building blocks, favor our preferred ingredients, small images
- Moving the broader **HPC community** forward
  - CUDA enabling 3rd-party runtimes and orchestration layers
  - Identifying and addressing technical challenges in the community

NGC: GPU-OPTIMIZED SOFTWARE HUB

Simplifying DL, ML and HPC Workflows

INDUSTRY SOLUTIONS

SMART CITIES		MEDICAL IMAGING
Parking Management	Traffic Analysis	Organ Segmentation
DeepStream SDK		Clara SDK

DEEP LEARNING MODEL SCRIPTS

Classification | Translation | Text to Speech | Recommender | . . .

50+ Containers
DL | ML | HPC

35 Models

Simplify Deployments

Innovate Faster

Deploy Anywhere

THE DESTINATION FOR GPU-OPTIMIZED SOFTWARE

HPC	Deep Learning	Machine Learning	Inference	Visualization	Infrastructure
BigDFT	Caffe2	Dotscience	DeepStream	CUDA GL	Kubernetes on NVIDIA GPUs
CANDLE	Chainer	H2O Driverless AI	DeepStream 360d	Index*	
CHROMA*	CT Organ Segmentation	Kinetica	TensorRT	ParaView*	
GAMESS*	CUDA	MapR	TensorRT Inference Server	ParaView Holodeck	
GROMACS	Deep Cognition Studio	MATLAB		ParaView Index*	
HOOMD-blue*	DeepStream 360d	OmniSci (MapD)		ParaView Optix*	
LAMMPS*	DIGITS	RAPIDS		Render server	
Lattice Microbes	Kaldi			VMD*	
Microvolution	Microsoft Cognitive Toolkit				
MILC*	MXNet				
NAMD*	NVCaffe				
Parabricks	PaddlePaddle				
PGI Compilers	PyTorch				
PIConGPU*	TensorFlow*				
QMCPACK*	Theano				
RELION	Torch				
	TLT Stream Analytics IVA				

*Multi-node HPC containers
New since SC18

NGC registration not required as of Nov'18

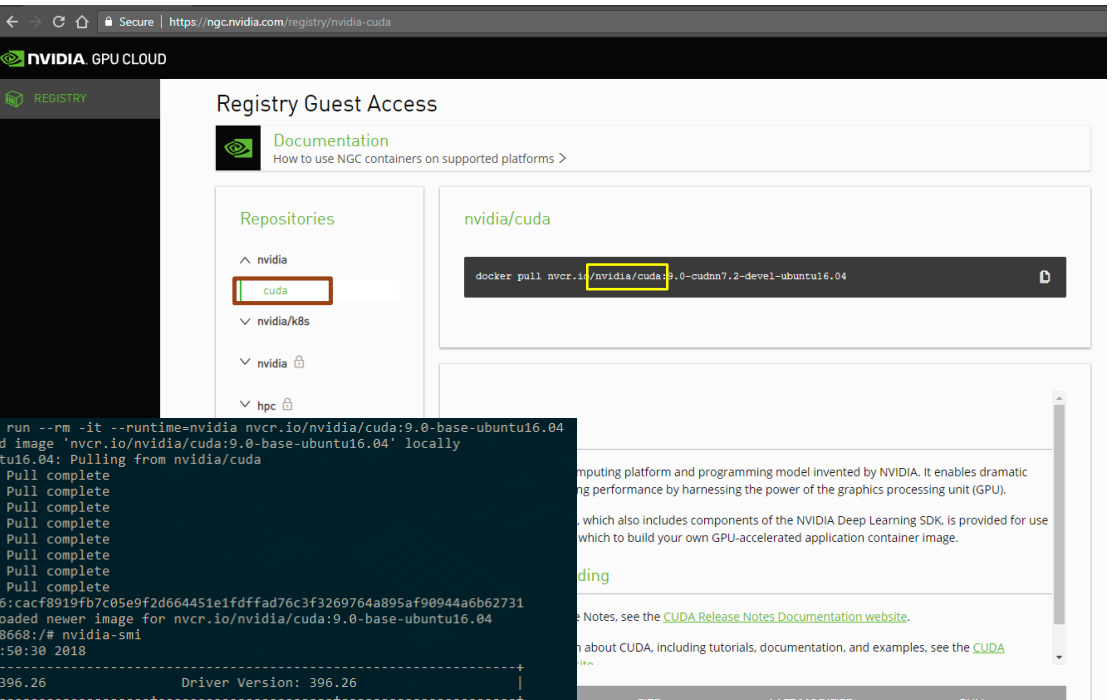
10 containers — October 2017

SOFTWARE ON THE NGC CONTAINER REGISTRY

48 containers — ~March 2019

CUDA CONTAINERS ON NVIDIA GPU CLOUD

- CUDA containers available from NGC Registry at **nvcr.io/nvidia/cuda**
- Three different flavors:
- **Base**
  - Contains the minimum components required to run CUDA applications
- **Runtime**
  - Contains *base* + CUDA libraries (e.g. cuBLAS, cuFFT)
- **Devel**
  - Contains *runtime* + CUDA command line developer tools. Some *devel* tags also include cuDNN




```
$ sudo docker run --rm -it --runtime=nvidia nvcr.io/nvidia/cuda:9.0-base-ubuntu16.04
Unable to find image 'nvcr.io/nvidia/cuda:9.0-base-ubuntu16.04' locally
9.0-base-ubuntu16.04: Pulling from nvidia/cuda
b234f539f7a1: Pull complete
55172d420b43: Pull complete
5ba5bbeb6b91: Pull complete
43ae2841ad7a: Pull complete
f6c9c6de4190: Pull complete
d5ddb2a2159f: Pull complete
663648e540ff: Pull complete
d056eaf3dff4: Pull complete
Digest: sha256:cacf8919fb7c05e9f2d664451e1fdffad76c3f3269764a895af90944a6b62731
Status: Downloaded newer image for nvcr.io/nvidia/cuda:9.0-base-ubuntu16.04
root@57dc39698668:/# nvidia-smi
Tue Sep 11 23:50:30 2018
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 396.26                 Driver Version: 396.26                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  GeForce GT 710      Off  | 00000000:01:00.0 N/A |                  N/A |
| 40%   36C    P8    N/A /  N/A |    396MiB /  2000MiB |     N/A      Default |
+-------------------------------+----------------------+----------------------+
|   1  TITAN V             Off  | 00000000:02:00.0 Off |                  N/A |
| 28%   42C    P8    27W / 250W |      0MiB / 12066MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0                    Not Supported                                       |
+-----------------------------------------------------------------------------+
root@57dc39698668:/#
```