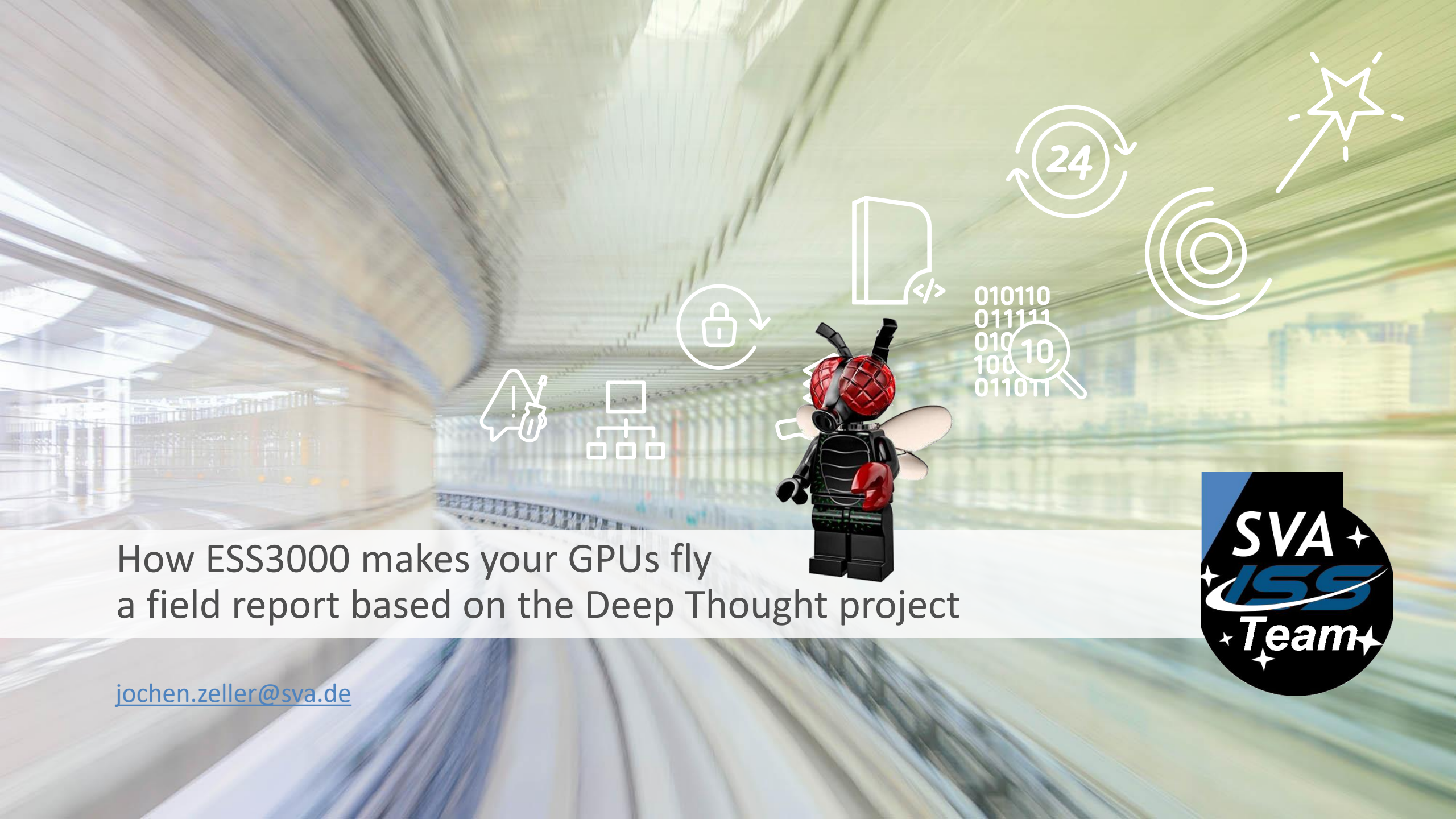
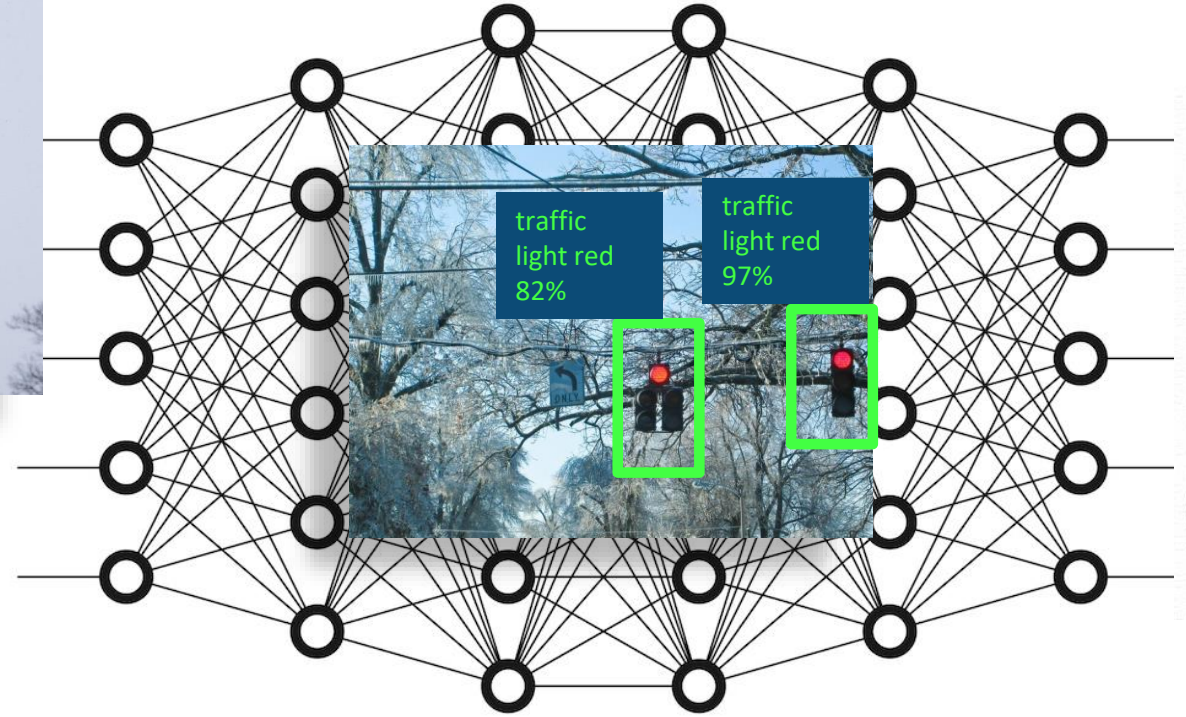


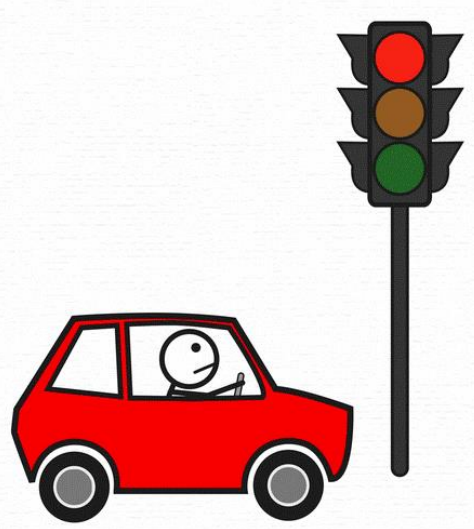
How ESS3000 makes your GPUs fly
a field report based on the Deep Thought project

jochen.zeller@sva.de

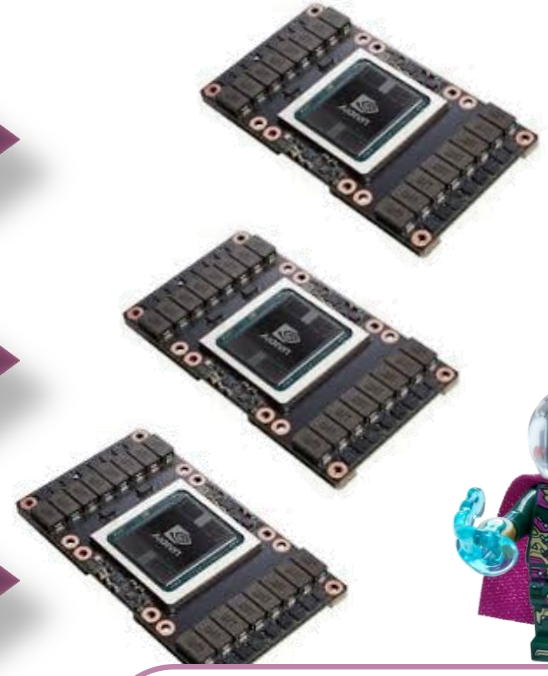
/ DEEP THOUGHT – AN AI PROJECT

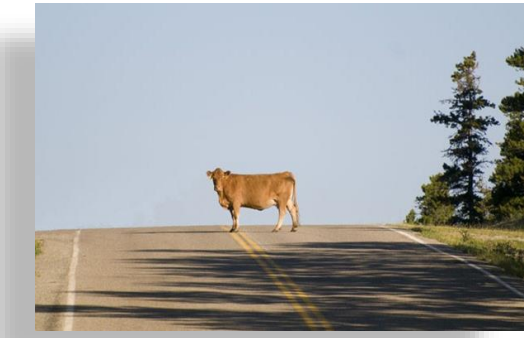
Neuronal network to detect traffic related situations





/ WHY DO THEY NEED SPECTRUM SCALE AND ESS3000





Many GPUs in a lot of DGX‘s need access to the images, in parallel and very fast without wasting GPU time with waiting for images.

/ WHY DO THEY NEED SPECTRUM SCALE AND ESS3000

Slightly more detailed:

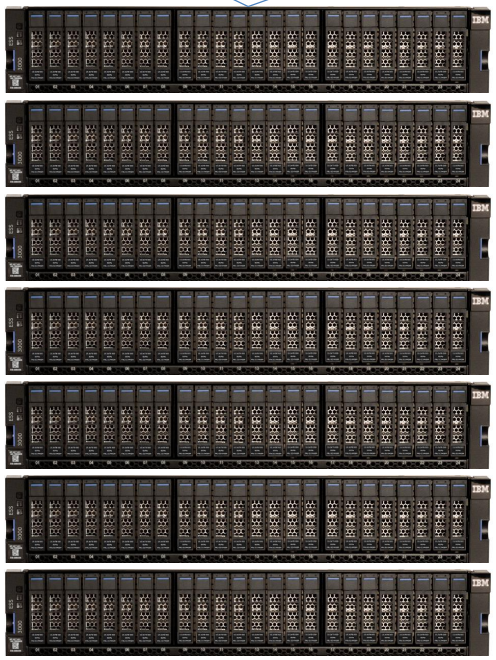
- hundreds of millions of images
- hundreds of GPU
- up to 500,000 images per epoche / learning cycle
- Important: reduce time to load images from storage / file system
  - Typical disk: read 500,000 images with a disk latency of 2 ms:
    - Synchronous read: ~ 16 minutes @ ~ 260MB/sec
    - Parallel read with 16 threads: ~ 1 minute @ ~ 4,100MB/sec
  - NVMe / fast disk: Read 500,000 images with a disk latency of 0.2 ms:
    - Synchronous read: ~ 100 seconds @ ~ 2,500MB/sec
    - Parallel read with 16 threads: ~ 7 seconds @ ~ ***40,000MB/sec***

/ DEEP THOUGHT INFRASTRUCTURE

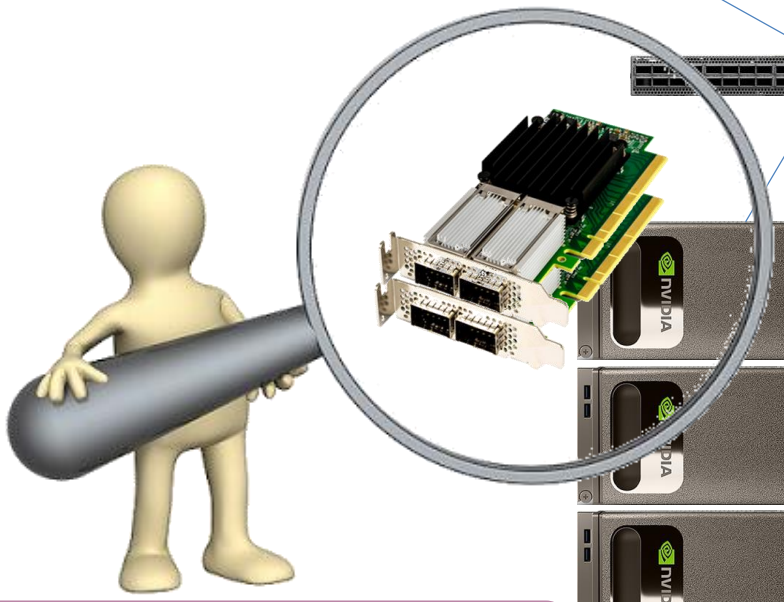
A total of 22 HDR switches
(6 spines and 16 leafs)
Non blocking!

Infiniband HDR

Infiniband EDR



7 x ESS3000



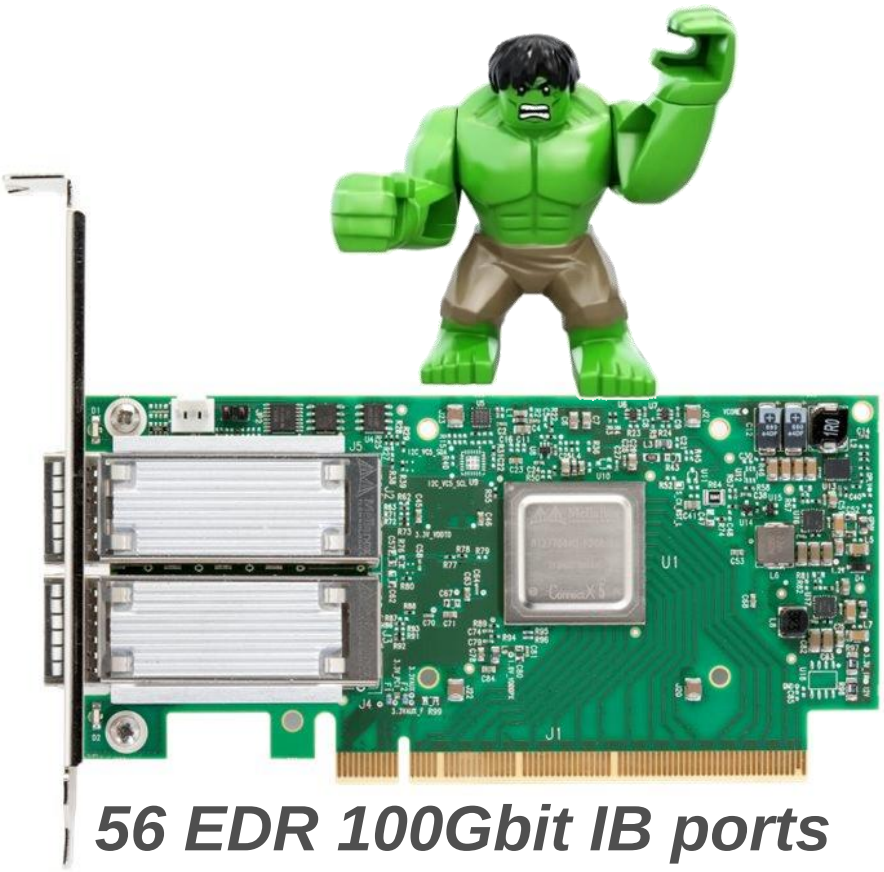
4 x EDR 100GB per DGX-1
8 x EDR 100GB per DGX-2

Infiniband EDR

many DGX-1 and DGX2

/ ESS3000 SPECS @ DEEP THOUGHT

- 7 x ESS3k / each with 2 IO nodes
  Per ESS3k:
    24 x 1,92TB NVMe
  Per IO node (node canister):
    2 x 14 core Intel x86_64
    768GB Memory
    2 x 2-port EDR 100GBit Infiniband
    RHEL8



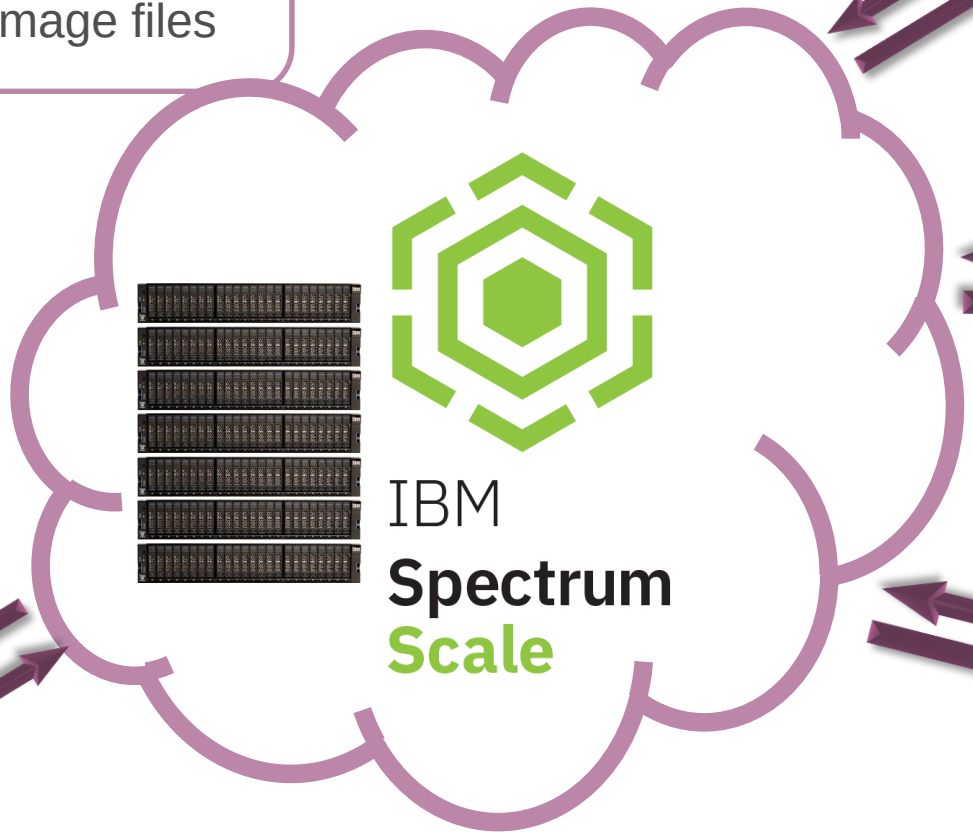
56 EDR 100Gbit IB ports

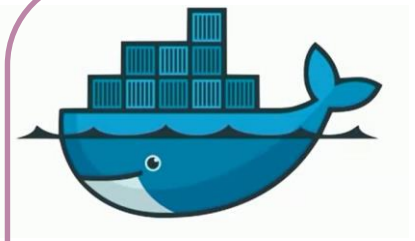
/ DEEP THOUGHT ECOSYSTEM

Spectrum Scale immutability to protect image files

MINIO

S3 / Object data transfer





Kubernetes managed container environment to run AI stack on Spectrum Scale

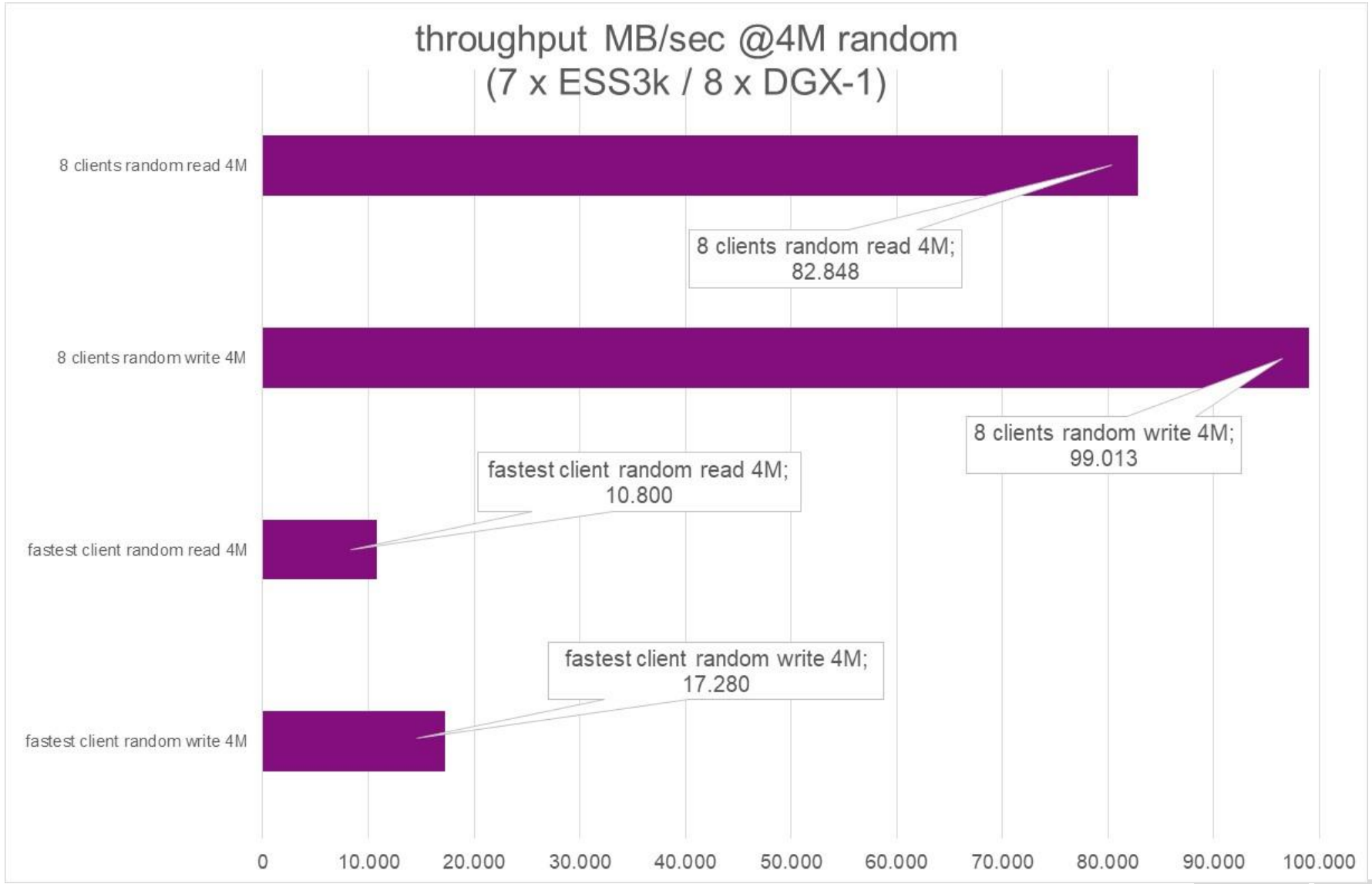
Workstation access by CES SMB

TCT based file copy to AWS

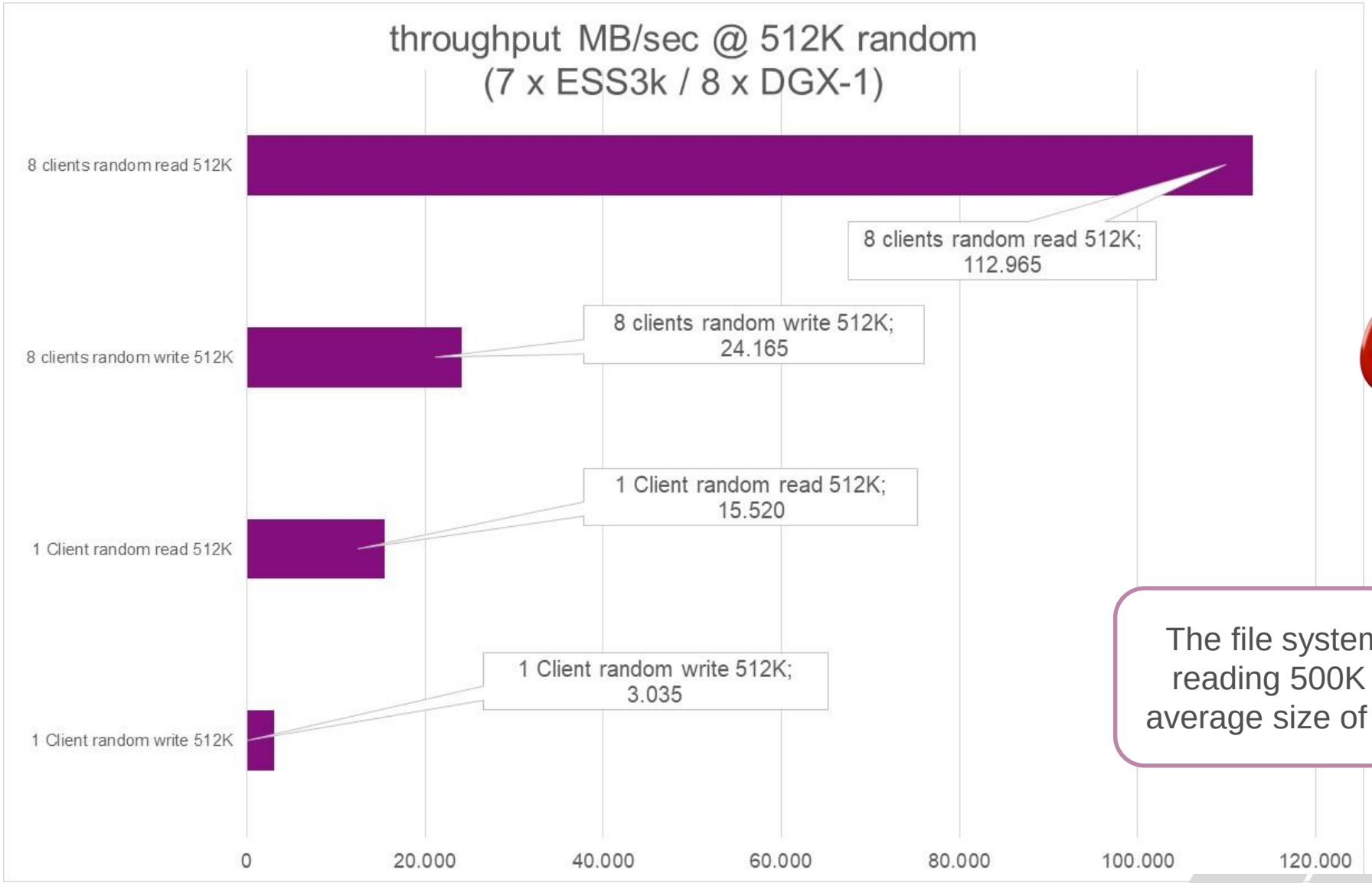
/ DEEP THOUGHT ESS3000 SETUP

- One recovery group per ESS3k chassis (new shared RG design)
- four vdisks per chassis
- two vdisks hosted per node canister
- 28 vdisks, all data and metadata in system pool
- One file system, 4MB block size

/ 8 DGX-1 RUMBLES ON THE FILE SYSTEM – 4MB IO SIZE

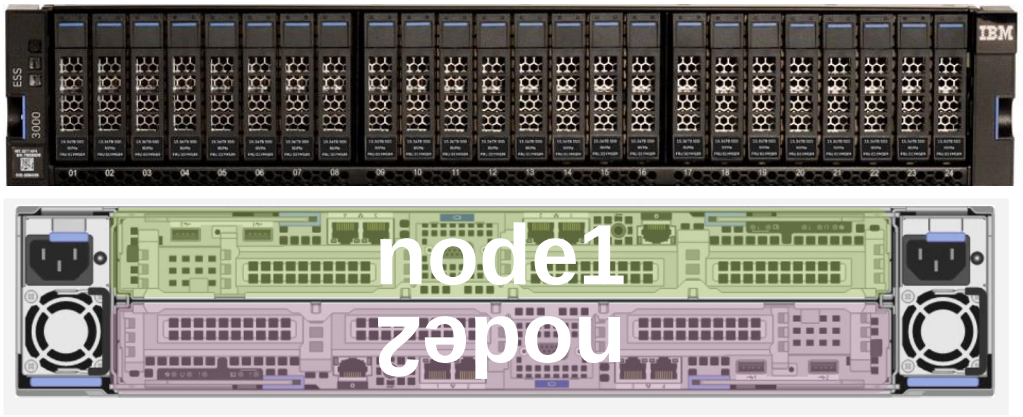


/ 8 DGX-1 RUMBLES ON THE FILE SYSTEM – 512K IO SIZE



The file system is optimized for reading 500K files – this is the average size of the traffic images.

/ ESS3000 – THAT'S AWESOME

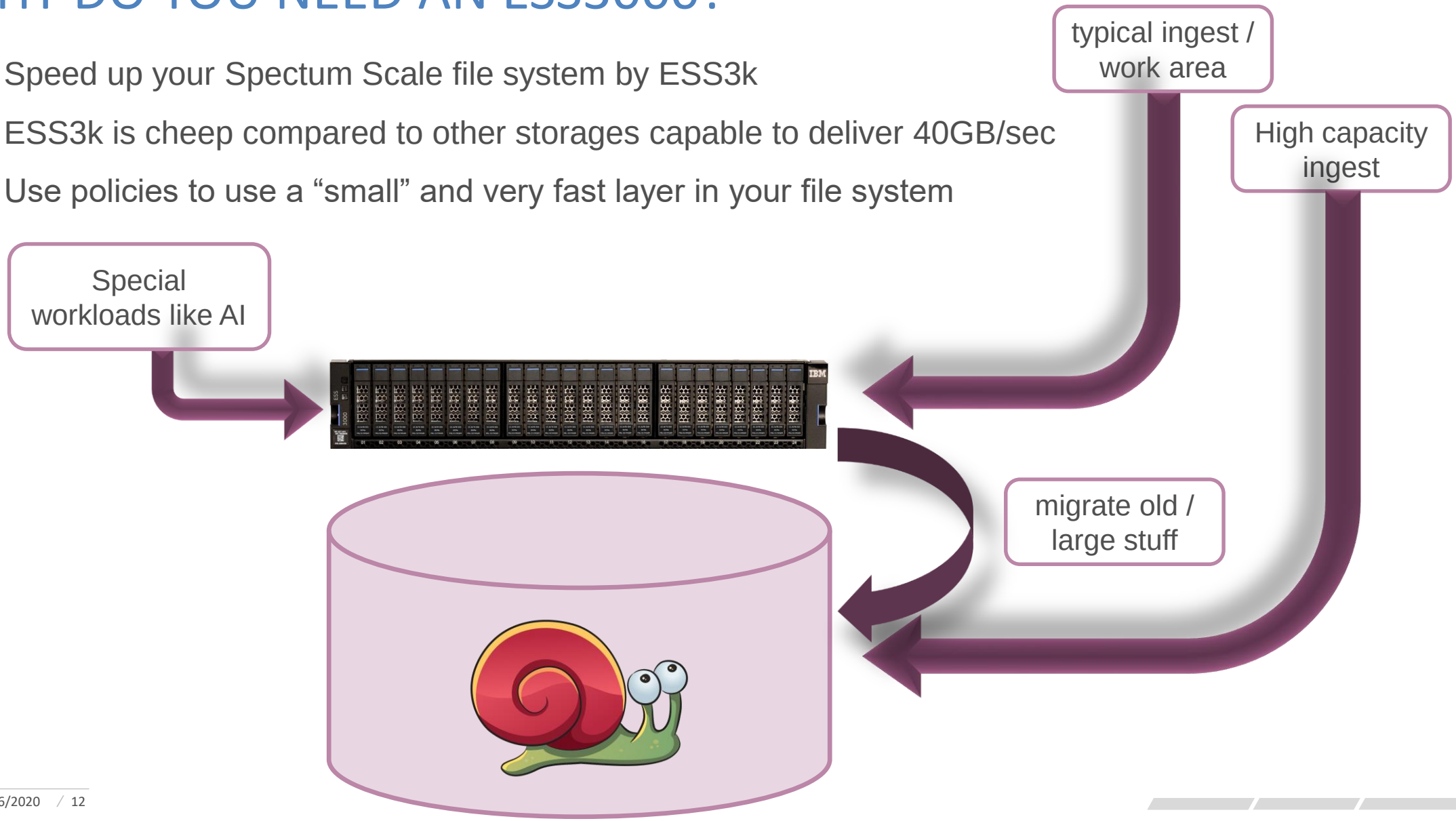


- Quiz: how many commands do you need to update 7 ESS3k?
  - Scale code
  - Native raid code
  - RHEL
  - ofed
  - Firmware (x86_64, adapter, NVMEs)

Just one: # ess3krun -G *ess_x86_64* **update**

/ WHY DO YOU NEED AN ESS3000?

- Speed up your Spectum Scale file system by ESS3k
- ESS3k is cheep compared to other storages capable to deliver 40GB/sec
- Use policies to use a “small” and very fast layer in your file system



/ WHY DO YOU NEED AN ESS3

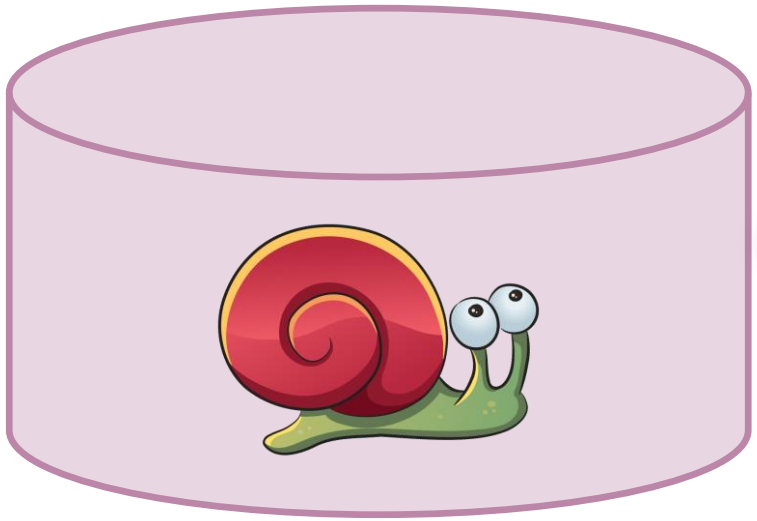
- Speed up your Spectum Scale file system
- ESS3k is cheep compared to other storage
- Use policies to use a “small” and very fast

Hausaufgabe für alle:

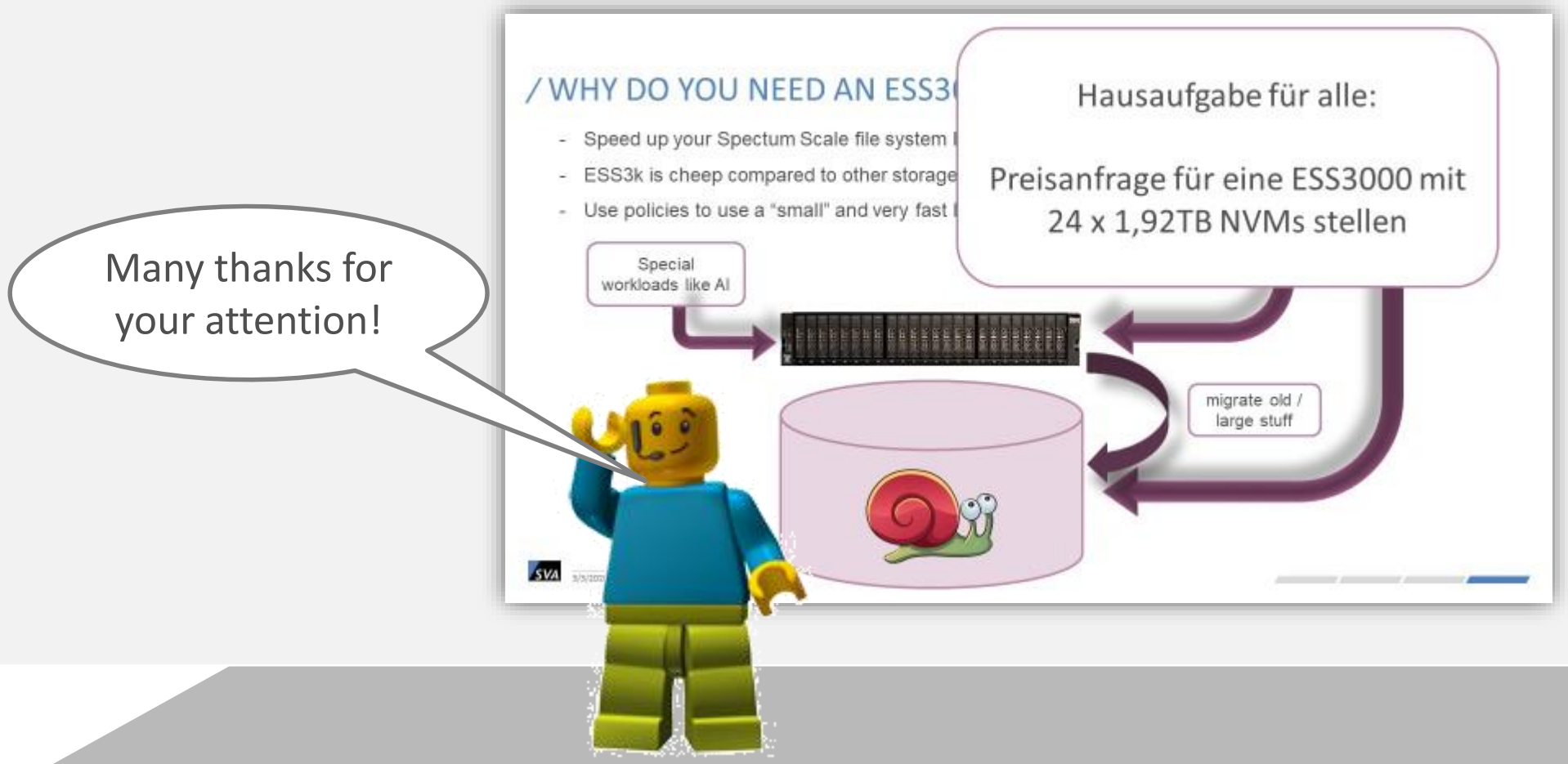
Preisanfrage für eine ESS3000 mit 24 x 1,92TB NVMs stellen

Special workloads like AI

migrate old / large stuff



/ THE END!


Many thanks for your attention!
/ WHY DO YOU NEED AN ESS3
- Speed up your Spectum Scale file system
- ESS3k is cheep compared to other storage
- Use policies to use a "small" and very fast
Special workloads like AI
Hausaufgabe für alle:
Preisanfrage für eine ESS3000 mit 24 x 1,92TB NVMs stellen
migrate old / large stuff

/ CONTACT

JOCHEN ZELLER

System Architekt

Tel.: +49 151 180 256 77

Mail: jochen.zeller@sva.de