Hewlett Packard
Enterprise

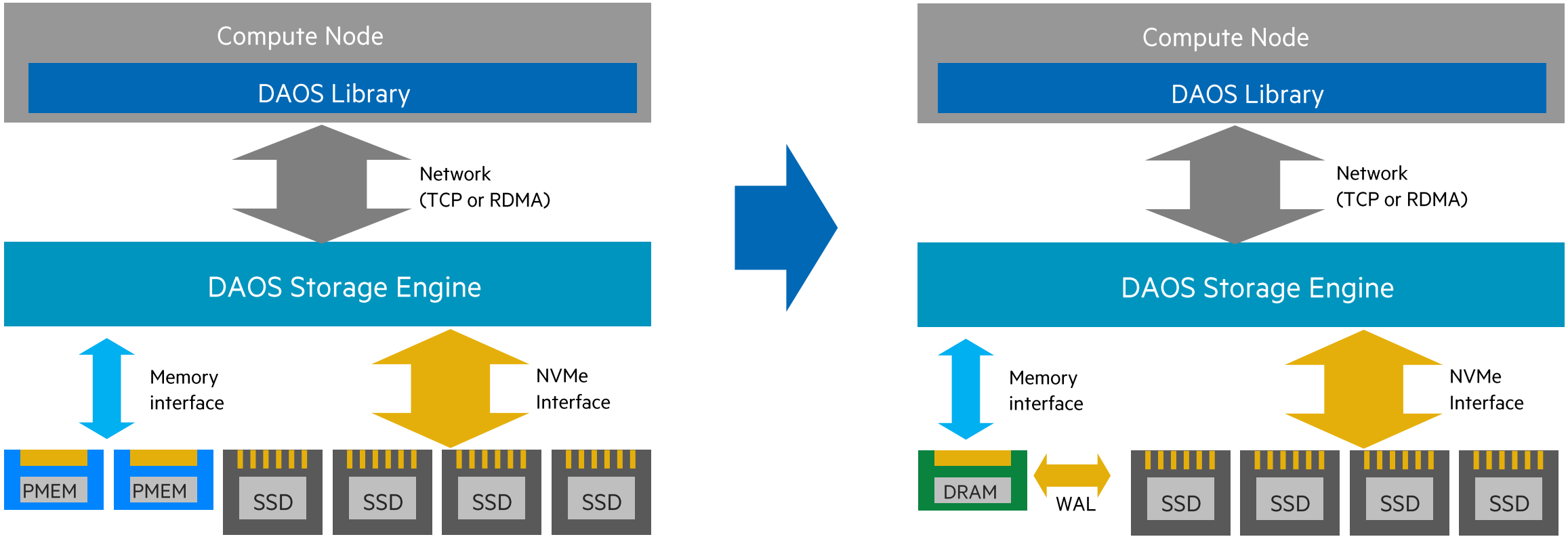
Virtual DAOS User Group (vDUG'26)

Pool creation with PMem, MD-on-SSD (Phase1, Phase2)

Michael Hennecke

21-May-2026

DAOS Architecture Evolution (with / without PMem)

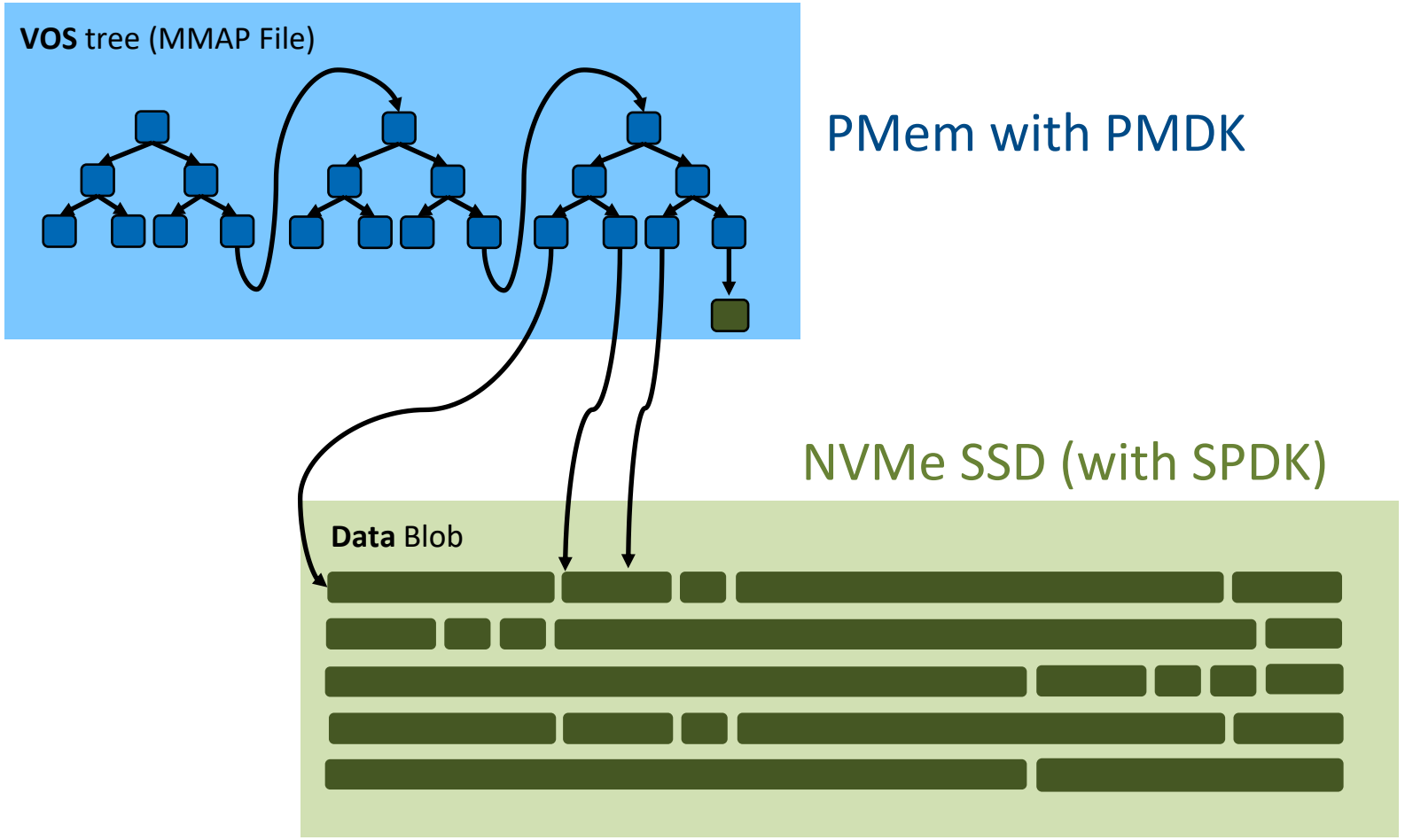


With Persistent Memory

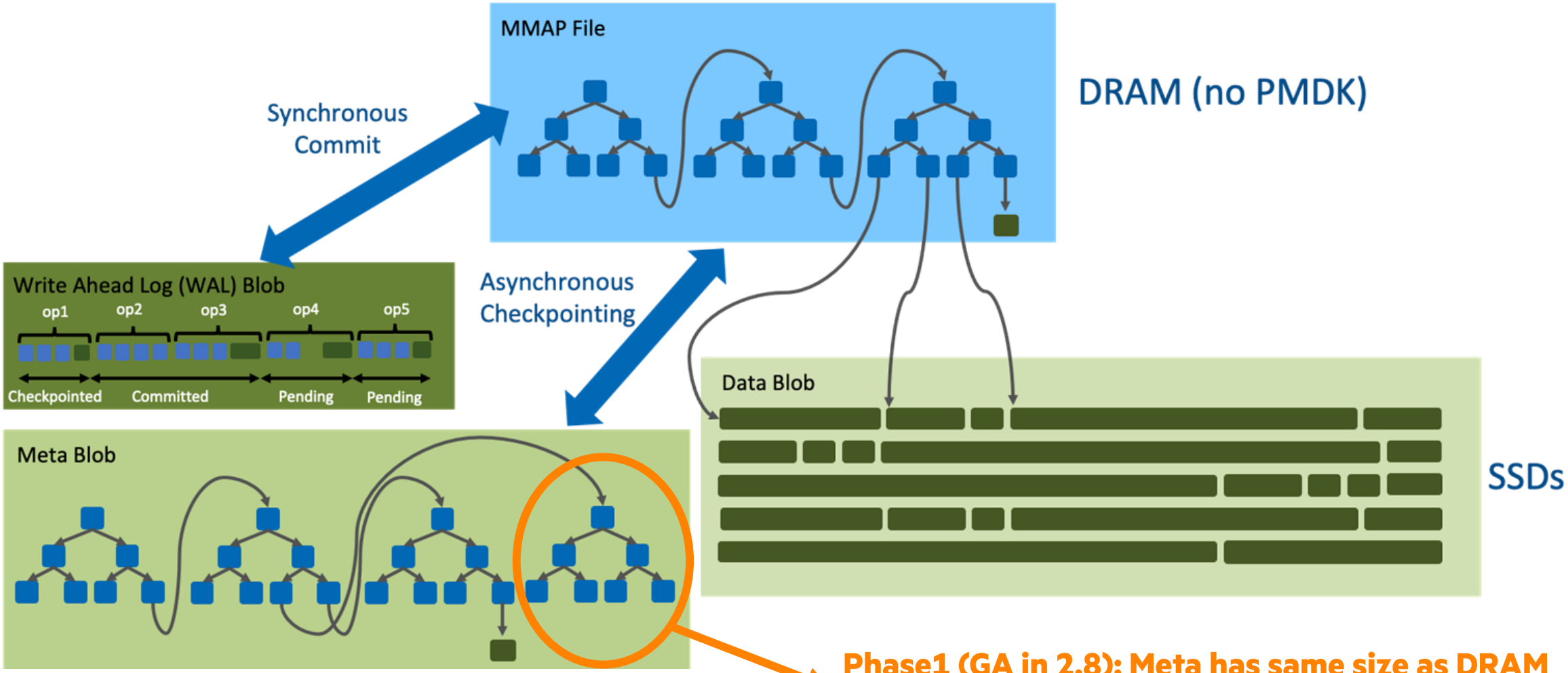
Without Persistent Memory

See the ISC23 workshop paper ...

Old DAOS Metadata Backend with PMem



New DAOS Metadata Backend without PMem



Phase1 (GA in 2.8): Meta has same size as DRAM

Phase2 (Tech Preview in 2.8): Meta can be > DRAM

Which one do I have? Configuration in daos_server.yml

```yaml
storage:
-
  class: dcpm
  scm_mount: /mnt/pmem1
  scm_list:
  - /dev/pmem1
-
  class: nvme
  bdev_list:
  - "0000:e3:00.0"
  - "0000:e4:00.0"
  - "0000:e5:00.0"
  - "0000:e6:00.0"
```

PMem-based DAOS

```yaml
storage:
-
  class: ram
  scm_mount: /mnt/dram1
  scm_size: 156

-
  class: nvme
  bdev_list:
  - "0000:e3:00.0"
  - "0000:e4:00.0"
  - "0000:e5:00.0"
  - "0000:e6:00.0"
```

"Ephemeral" DAOS

```yaml
storage:
-
  class: ram
  scm_mount: /mnt/dram1
  scm_size: 156
-
  class: nvme
  bdev_roles:
  - wal
  - meta
  - data
  bdev_list:
  - "0000:e3:00.0"
  - "0000:e4:00.0"
  - "0000:e5:00.0"
  - "0000:e6:00.0"
```

DAOS with **MD-on-SSD**

- no distinction here if Phase1 or Phase2...

See the ISC23 workshop paper ...

dmg pool create

```
[root@N0201 ~]# dmg pool create --help
Usage:
  dmg [OPTIONS] pool create [create-OPTIONS] [<pool label>]

[create command options]
  -g, --group=      DAOS pool to be owned by given group, format name@domain
  -u, --user=       DAOS pool to be owned by given user, format name@domain
  -P, --properties= Pool properties to be set
  -a, --acl-file=   Access Control List file path for DAOS pool
  -z,--size=   Total size of DAOS pool or its percentage ratio (auto)
  -t, --tier-ratio= Percentage of storage tiers for pool storage (auto; default: 6,94)
  -k, --nranks=     Number of ranks to use (auto)
  -v, --nsvc=       Number of pool service replicas
  -s, --scm-size=   Per-engine SCM allocation for DAOS pool (manual)
  -n, --nvme-size=  Per-engine NVMe allocation for DAOS pool (manual)
      --meta-size=  Per-engine Metadata-on-SSD allocation for DAOS pool (manual). Only valid in MD-on-SSD mode
      --data-size=  Per-engine Data-on-SSD allocation for DAOS pool (manual). Only valid in MD-on-SSD mode
      --mem-ratio=  Percentage of the pool metadata storage size (on SSD) that should be used as the
                    memory file size (on ram-disk). Default value is 100% and only valid in MD-on-SSD mode
  -r, --ranks=      Storage engine unique identifiers (ranks) for DAOS pool
```

Pool Sizing Gotchas: --size is (SCM + NVMe) … and beware of --tier-ratio

- DAOS software's default `--tier-ratio=6,94` means 6.00% of SCM and 94.00% of NVMe
  - Actual DAOS server hardware will often have very different ratios…

- Examples of PMem-based DAOS systems:
  - ALCF Aurora: 16x 512 GiB = 8 TiB PMem, 16x 15.36 TB NVMe → DRAM tier ratio is 8/(8+223) ~ 3.5 %
  - LRZ SNG2: 16x 128 GiB = 2 TiB PMem, 8x 3.84 TB NVMe → DRAM tier ratio is 2/(2+28) ~ 6.6 %
  - ZIB Lise: 12x 128 GiB = 1.5 TiB PMem, 4x 7.68 TB NVMe → DRAM tier ratio is 1.5/(1.5+28) ~ 5.1%

- Two gotchas for pool creation – true for both PMem and MD-on-SSD (Phase1):
  1. If a user wants a pool with 100TB (with default tier ratio), need to request `--size=106TB` (100TB / 94%)
     - Most relevant for PMem-based systems (MD-on-SSD systems have smaller DRAM capacity and more NVMe),
  2. If the hardware's SCM is smaller than 6%, default will leave **lots** of NVMe capacity "stranded" → use the actuals!
     - Aurora with default tier ratio would only allocate 8 TiB / 6% = 133 TiB, leaving ~40% (90 TiB) NVMe unusable (per node)
     - Very important to use the actual sizes for MD-on-SSD (Phase1), where actual tier ratios are typically <1%

- DAOS-18112 feature request to set --tier-ratio based on actual hardware (not committed to a release yet)

Hewlett Packard Enterprise | DAOS

1 TB = 0.909495 TiB …

DRAM Memory Ratios of the Supported HPE K3000 v1 Server Configurations

NVMe	NVMe size	DRAM	NVMe size	DRAM	NVMe size	DRAM
Qty	3,84		7,68		15,36	
8	30,72	256	61,44	512	122,88	1.024
	27,94	0,250	55,88	0,500	111,76	1,000
		0,887%		**0,887%**		**0,887%**
12	46,08	512	92,16	768	184,32	2.048
	41,91	0,500	83,82	0,750	167,64	2,000
		1,179%		**0,887%**		**1,179%**
16	61,44	512	122,88	1.024	245,76	2.048
	55,88	0,500	111,76	1,000	223,52	2,000
		0,887%		**0,887%**		**0,887%**
20	76,80	768	153,60	2.048	307,20	2.048
	69,85	0,750	139,70	2,000	279,40	2,000
		1,062%		**1,411%**		**0,711%**

Key:

NVMe TB	DRAM GiB
NVMe TiB	DRAM TiB
--	**DRAM ratio**

Base10 / Base2 Units:

tera (10^20)	1000000000000
tebi (2^40)	1099511627776
TiB = TB *	**0,909494702**

Hewlett Packard Enterprise | DAOS

What changes for MD-on-SSD Phase2 (Tech Preview in DAOS 2.8) ?

- For Phase1, meta blob size on NVMe is identical to size of VOS files in DRAM (`--mem-ratio 100%`)
- For Phase2, DRAM is still limited by server hardware, but meta blob size on NVMe can be larger
  - And the `--tier-ratio` in MD-on-SSD refers to the meta blob size on NVMe, not to the "SCM" DRAM size…

```
# dmg pool create -u hennecke -g luser --size 300T
  --tier-ratio 0.60,99.40 my_pool01

NOTICE: SCM:NVMe ratio is less than 1.00%, DAOS
performance will suffer!
Creating DAOS pool with automatic storage
allocation: 300 TB total, 0.60% ratio

Pool created with 0.60%,99.40% storage tier ratio
-----------------------------------------------
  UUID : 449ca480-e556-40ec-8b92-6be58c8c6eb5
  Service Leader    : 1
  Service Ranks     : [0-1]
  Storage Ranks     : [0-1]
  Total Size        : 300 TB
  Metadata Storage  : 1.8 TB (900 GB / rank)
  Data Storage      : 298 TB (149 TB / rank)
  Memory File Size  : 1.8 TB (900 GB / rank)
```

```
# dmg pool create -u hennecke -g luser --size 300T
  --tier-ratio 6.00,94.00 --mem-ratio 10% my_pool01


Creating DAOS pool with automatic storage
allocation: 300 TB total, 6.38% ratio

Pool created with 6.00%,94.00% storage tier ratio
-----------------------------------------------
  UUID : 011e2677-d1d6-466b-8b66-6a0dad52c97a
  Service Leader    : 1
  Service Ranks     : [0-1]
  Storage Ranks     : [0-1]
  Total Size        : 300 TB
  Metadata Storage  : 18 TB (9.0 TB / rank)
  Data Storage      : 282 TB (141 TB / rank)
  Memory File Size  : 1.8 TB (900 GB / rank)
```

Thank you

michael.hennecke@hpe.com

Hewlett Packard Enterprise | DAOS