TOSHIBA

实验报告

PROMISE VTrak J5960 JBOD

前言

气候中和与可持续发展是现代数据中心關心的重要議題。目标：降低功耗和总体碳足迹，支持全天候处理主要存储在硬盘驱动器上的海量数据。

绿色倡议影响数据中心的各个方面，从余热再利用技术和自然冷却技术，到采用可持续生产的库存和先进的硬盘驱动器 (HDD) 技术，都需要纳入考虑。

以下是实验室报告内容：

PROMISE 的新型 VTrak J5960 4U 60 盘位顶部装载式 JBOD 的宣传标语是“拥有绿色 DNA 的 JBOD”，这句标语包含了对环境保护和可持续生产的承诺。东芝电子欧洲有限公司（“东芝”）利用机会在实验室测评了这款 JBOD，测评采用 60 个数据存储容量为 18TB 的东芝企业级硬盘，总容量高达 1080TB。

东芝对这款 JBOD 的功能、性能、噪音和功耗等方面进行了全面测评，重点评估了 PROMISE 声称的环保特性。

图 1：东芝实验室中的 PROMISE VTrak J5960 JBOD。

TOSHIBA

物理尺寸和机械特征：

J5960 的 4U 机箱长度只有 666 毫米，是我们在东芝硬盘实验室中见过最短的高密度 JBOD。它与常见的 66 厘米 2U 服务器机箱的长度一样，可以很方便地适配现有机架。与其他 JBOD 相比，这是一个很大优势：许多 JBOD 长度超过 1000 毫米，需要机架更长，否则容易产生布线问题。

J5960 的热插拔 IO 模块 (IOM) 可以拉出到 JBOD 的正面，而连接的电缆则保留在设备的背面。这样的设计使得在现场更换 IOM 非常容易，而常见的后部插拔设计则需要处理一堆电源线和信号线。

JBOD 的盖子设计非常特别。它的盖子可以连接机架，当 JBOD 被拉出进行维护时，盖子会留在机架上面。我们对此进行了测试，使用体验非常顺畅。因为无需抬起盖子，只需将 JBOD 拉出到故障硬盘的位置即可。它也可以安装在机架的顶部位置。

每个硬盘都用 4 颗螺丝固定在金属托盘中。

LED 状态指示灯通常处于关闭状态，只有当盖子被移除（即从机架上拉出 JBOD）时才会激活。这种设计方式可以节省几瓦的额外功率。

东芝实验室的设置

型号：	PROMISE VTrak J5960 4U-SAS-60-D BP
固件：	1023
主机操作系统：	Linux (Centos 7.9)
主机操作系统：	Windows (Windows Server 2019 Standard)
主机总线适配器 (HBA)：	Broadcom Avago HBA 9500-16e (Host IF: 8x PCIe-Gen4)
RAID 适配器：	Microchip Adaptec® SmartRAID Ultra 3254-16e/e (16x PCIe-Gen4)

使用企业级容量硬盘(SAS)进行测试：

型号名称:	Toshiba MG09SCA18TE
缓冲区大小:	512B emulated
固件:	0104

图2: 盖子打开状态的 J5960.

数据速率: 282MB/s

功耗

Idle_B:	3.36W
顺序写入:	7.62W
顺序读取:	8.71W
随机写入:	6.64W
随机读取:	9.47W

JBOD 基本功能:

基本功能:	ok
SAS IOM 检测:	ok
热插拔/重新插入:	ok
智能读取:	ok
箱柜管理:	ok, 经过 RJ11 串行连接进行验证

图 3：在 PROMISE 托盘中的 Toshiba MG09SCA18TE 硬盘。

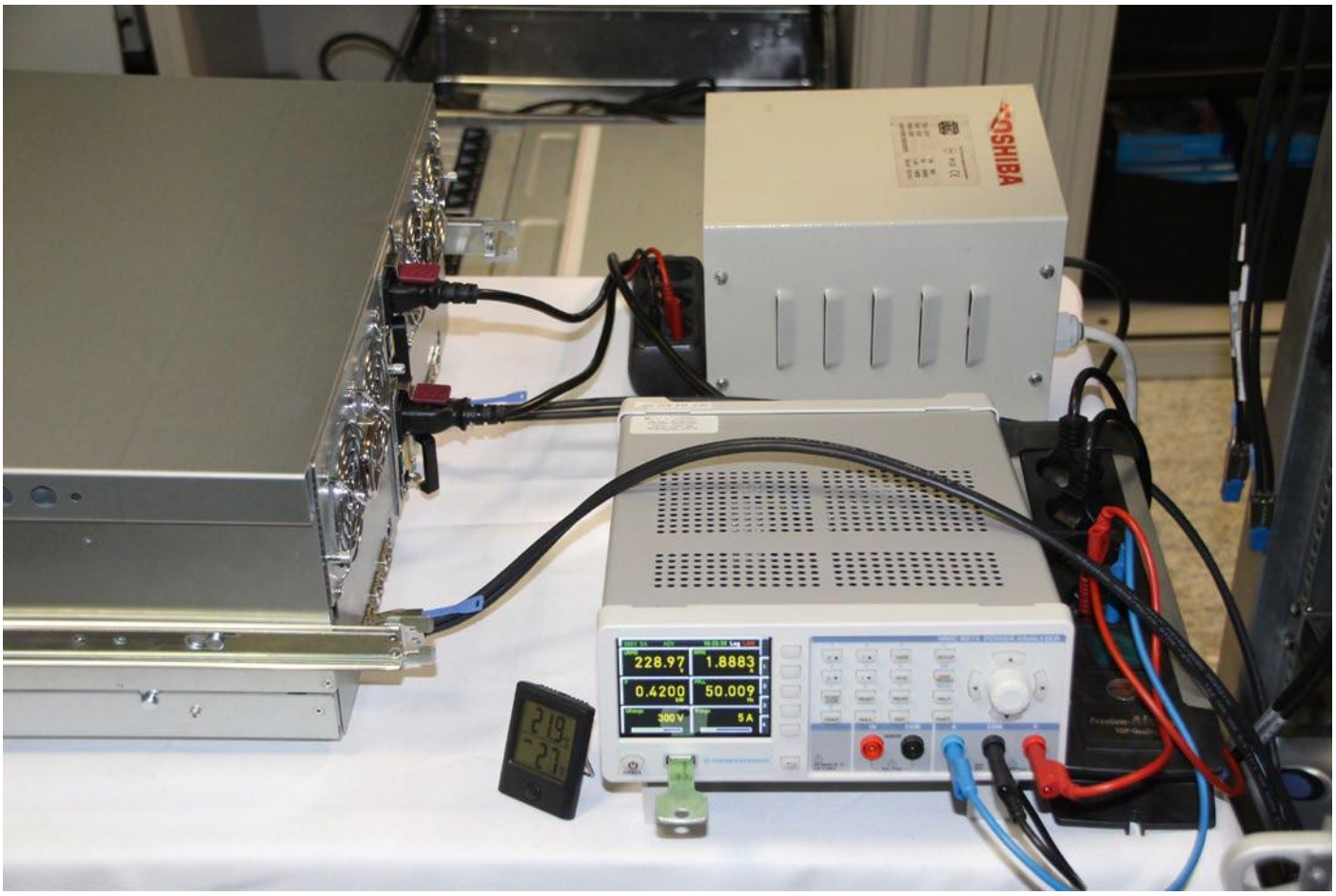

图 4：东芝硬盘实验室中的功率测量设置。

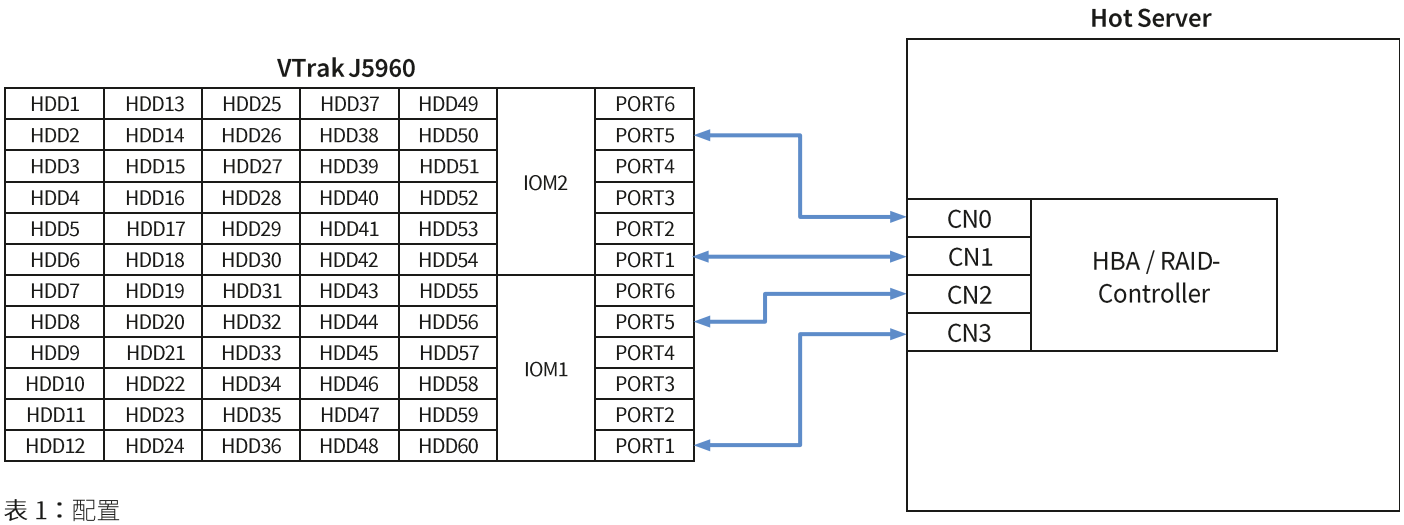

表 1：配置

为了精确测量功耗，我们使用了高精度的专业功率分析仪 (R&S HMC8015)。

JBOD on, no drives, SAS link to host on: 100W
JBOD with drives, maximum start-up power over 500ms..... 850W
JBOD with raw drives at HBA Idle_B:..... 305W
Lambda (ratio of active and reactive power).....0.96
Noise at 1 m distance 80dB
Temperature ambient..... 23°C

VTrak J5960 的默认设置是：如果 2 分钟左右没有被访问，则所有 (SAS) 硬盘都将被置于空闲状态。对于不带硬盘的双 IOM JBOD，100W 左右的功率值表现非常出色。在相同条件下，专用的单 IOM JBOD 的功率可低至 80W，但是双 IOM JBOD 的功率通常在 200-300W 之间。

“带硬盘且处于空闲状态”的功率只有 300W 左右，这个数值也非常低，其他的 60 盘位 JBOD 的功率通常不低于 400W。λ 系数高达 0.96，这意味着电源装置产生的无功功率比例非常低（相当于总功率的 4%）。λ 系数越高（越好），则无功功率越小。无功功率不需要成本并且不产生热量，但是，供电轨的尺寸必须同时考虑有功功率和无功功率——因此对于大型数据中心而言，高 λ 系数是一个重要考虑因素。

东芝实验室的性能测量

为了进行 SAS 硬盘测试，我们使用了两根 mini-SAS-HD 电缆，将 JBOD 的两个 IOM 分别连接到 16e HBA 和 RAID 控制器的 4 个 mini-SAS-HD 端口。

这种配置的理论 JBOD/HDD 访问带宽是 4x 4.8GB/秒 = 19.2GB/秒，但是需要通过安装多路径功能进行路径聚合。如需配置 HBA，必须在 Linux/Windows 中手动启用多路径。RAID 控制器（例如 Microchip Adaptec® Ultra-3254 型号）可以自动检测配置并调用正确的多路径设置。手动多路径和 SAS 链路聚合仅适合 SAS 硬盘。配置 SATA 硬盘仍需聚合 60 个硬盘的 IOPS，但顺序带宽通常受限于一个 mini-SAS-HD 链路（4.8GB/秒）。

我们使用“fio”（灵活的 IO 测试仪软件）测试了几种硬盘配置，分别测量了在顺序、随机和混合工作负载情况下的性能及功耗。我们测试了通过 HBA 连接、在 RAID 配置中作为实体硬盘和逻辑硬盘的单个硬盘。针对逻辑硬盘，我们还测量了复制（读取和写入）大文件的性能和功率。

JBOD 设置说明

VTrak J5960 的默认设置是“硬盘不活动超过 2 分钟则切换至空闲状态”。硬盘从空闲状态切换至活动状态大约需要 1200 毫秒。

使用 RAID 配置时，建议禁用此功能——因为在大型 RAID 中，即使 RAID 处于满数据负载的状态，部分硬盘不活动的时间也可能超过 2 分钟。如果 RAID 中的部分硬盘被切换至空闲模式，则访问这些硬盘时将会出现 1200 毫秒的延时。

禁用空闲模式切换：通过串行电缆连接到 IOM 管理端口 (115200/8/N/1)，CLI 命令为“enclosure -m -idlep 0”。

东芝执行测试的模式是“分区模式 0”（= 默认设置）。这表示“不分区”，因此所有硬盘都可以从两个 IOM 进行访问。

某些 RAID 控制器或 HBA 可能不支持 4 根电缆连接至两个 IOM。如果出现这种情况，可以使用“分区模式 1”进行变通。在该模式下，30 个硬盘从一个 SAS 端口连接至 IOM1，而另外 30 个硬盘则从一个 SAS 端口连接至 IOM2。这种配置方式相当于两个支持 30 个硬盘的 JBOD。更改分区模式的 CLI 命令为“enclosure -m -z 1”。

所有硬盘并行作为单个物理设备(多路径):

操作系统: Linux (Centos 7.9)
HBA/控制器: Broadcom HBA9500-16e
硬盘: 60x Toshiba MG09SCA18TE
配置: Dual IOM 2x 2 Mini-SAS HD cables (3m in length)
Multipath setup for disks

工作负载	功率(W)	IOPS	带宽(MB/s)
顺序写入 1024K	610		13300
顺序读取 1024K	640		14500
随机写入 4K	510	24100	
随机读取 4K	540	33900	
混合 4K/64K/256K/2M	540	22600	2350
环境温度	23°C		
最低温度硬盘	27°C		
最高温度硬盘	36°C		

理论最大带宽为282MB/秒（单盘）x 60 = 16.2GB/秒。14.5GB/秒及 20K+ IOPS 的配置接近理论极限。

最大 640W 的功率数据证明了 JBOD 的环保性能。驱动器之间的极限温差小于 10°C，而最高温度比环境温度低 14°C，可实现高效冷却，保证硬盘的可靠性和较长使用寿命。

所有硬盘作为 RAID10，Windows：

操作系统: Windows Server 2019
RAID/适配器: Microchip Adaptec® SmartRAID Ultra 3254-16e/e (16x PCIe-Gen4)
硬盘: 60x Toshiba MG09SCA18TE
配置: Dual IOM 2x 2 Mini-SAS HD cables (3m in length)

工作负载	功率(W)	IOPS	带宽(MB/s)
顺序写入 1024K	510		8600
顺序读取 1024K	570		15300
随机写入 4K	600	12800	
随机读取 4K	730	9900	
混合 4K/64K/256K/2M	680	6100	1800
空闲(RAID 后台)	470		
环境温度	25°C		
最低温度硬盘	29°C		
最高温度硬盘	38°C		

脚本 1 – 所有硬盘并行作为单个物理设备(多路径):

```
fio --direct=1 --bs=1m --iodepth=16 --size=32g --ioengine=libaio --group_reporting --rw=write --output=se-
qwrite.log --name=/dev/mapper/mpath{a..z} -- name=/dev/mapper/mpatha{a..z} --name=/dev/mapper/mpathb{a..h}

fio --direct=1 --bs=1m --iodepth=16 --size=32g --ioengine=libaio --group_reporting --rw=read --output=se-
qread.log --name=/dev/mapper/mpath{a..z} -- name=/dev/mapper/mpatha{a..z} --name=/dev/mapper/mpathb{a..h}

fio --direct=1 --bs=4k --iodepth=16 --size=512m --ioengine=libaio --group_reporting --rw=randwrite --out-
put=randwrite.log --name=/dev/mapper/mpath{a..z} -- name=/dev/mapper/mpatha{a..z} --name=/dev/mapper/mpath-
b{a..h}

fio --direct=1 --bs=4k --iodepth=16 --size=512m --ioengine=libaio --group_reporting --rw=randread --out-
put=randread.log --name=/dev/mapper/mpath{a..z} -- name=/dev/mapper/mpatha{a..z} --name=/dev/mapper/mpath-
b{a..h}

fio --direct=1 --bssplit=4k/20:64k/50:256k/20:2M/10 --iodepth=16 --size=8g -ioengine=libaio --group_report-
ing --rw=randrw --output=mixed.log --name=/dev/mapper/mpath{a..z} -- name=/dev/mapper/mpatha{a..z} --name=/
dev/mapper/mpathb{a..h}
```

脚本 2 – 所有硬盘作为 RAID10，Windows 物理硬盘:

```
fio --filename=\\.\Physicaldrive1 --direct=1 --rw=write --bs=1m --iodepth=16 --time_based --runtime=300
--group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64 --norandommap --randrepeat=0
--output=seqwritephysical.log

fio --filename=\\.\Physicaldrive1 --direct=1 --rw=read --bs=1m --iodepth=16 --time_based --runtime=300
--group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64 --norandommap --randrepeat=0
--output=seqreadphysical.log

fio --filename=\\.\Physicaldrive1 --direct=1 --rw=randwrite --bs=4k --iodepth=16 --time_based --runtime=300
--group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64 --norandommap --randrepeat=0
--output=randwritephysical.log

fio --filename=\\.\Physicaldrive1 --direct=1 --rw=randread --bs=4k --iodepth=16 --time_based --runtime=300
--group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64 --norandommap --randrepeat=0
--output=randreadphysical.log

fio --filename=\\.\Physicaldrive1 --direct=1 --rw=randrw --bssplit=4k/20:64k/50:256k/20:2M/10 --iodepth=16
--time_based --runtime=300 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64
--norandommap --randrepeat=0 --output=mixedphysical.log
```

由于 RAID10 中的写入始终分为并行的两个镜像设备，因此写入速度会减半。RAID 配置中的空闲功率几乎相当于有功功率，因为 RAID 控制器会不停访问驱动器进行一些后台一致性检查。

工作负载	功率(W)	IOPS	带宽(MB/s)
顺序写入 1024K	520		6900
顺序读取 1024K	550		15000
随机写入 4K	520	11100	
随机读取 4K	540	29500	
混合 4K/64K/256K/2M	550	8100	2400
Windows 副本	500		550
空闲(RAID 后台)	470		
环境温度	25°C		
最低温度硬盘	29°C		
最高温度硬盘	39°C		

所有硬盘作为 RAID10，Windows 逻辑卷:

操作系统: Windows Server 2019
RAID/适配器: Microchip Adaptec® SmartRAID Ultra 3254-16e/e (16x PCIe-Gen4)
硬盘: 60x Toshiba MG09SCA18TE
配置: Dual IOM 2x 2 Mini-SAS HD cables (3m in length)

脚本 3 – 所有硬盘作为 RAID10，Windows 逻辑卷:

```
fio --filename=test --size=1T --direct=1 --rw=write --bs=1m --iodepth=16 --time_based --runtime=300
--group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64 --norandommap --randrepeat=0
--output=seqwritelogical.log

fio --filename=test --size=1T --direct=1 --rw=read --bs=1m --iodepth=16 --time_based --runtime=300 --group_
reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64 --norandommap --randrepeat=0 --out-
put=seqreadlogical.log

fio --filename=test --size=1T --direct=1 --rw=randwrite --bs=4k --iodepth=16 --time_based --runtime=300
--group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64 --norandommap --randrepeat=0
--output=randwritelogical.log

fio --filename=test --size=1T --direct=1 --rw=randread --bs=4k --iodepth=16 --time_based --runtime=300
--group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64 --norandommap --randrepeat=0
--output=randreadlogical.log

fio --filename=test --size=1T --direct=1 --rw=randrw --bssplit=4k/20:64k/50:256k/20:2M/10 --iodepth=16
--time_based --runtime=300 --group_reporting --name=job1 --ioengine=windowsaio --thread --numjobs=64
--norandommap --randrepeat=0 --output=mixedlogical.log
```

针对 Windows 逻辑卷的基准测试并未使用整个驱动器大小，而是使用了 1TB 大小的测试文件进行操作。因为这样更符合实际用例。IOPS 超过 32k 是因为寻道操作没有覆盖全部容量范围。与之前对物理驱动器进行 500TB 的完整寻道相比，这种操作方式可以显著降低写入操作的功耗。

配置 SATA硬盘

我们还使用 SATA 硬盘（型号 MG09ACA18TE）对 VTrak J5960 JBOD 进行了测试。由于 SATA 接口只有一个信号路径，我们使用了单 IOM 配置（IOM2 被拔出，四根电缆全部连接到 IOM1）。

顺序性能受限于一根 mini-SAS HD 电缆的水平（顺序读写约为 4.3GB/秒）。IOPS 值与基于 SAS 的测试结果几乎相同，因为它们不受带宽限制（完整容量为 10k IOPS，具有 1TB 数据范围的逻辑硬盘为 30k IOPS）。

在同等设置条件下，使用 SATA 硬盘进行单 IOM 操作的功耗比使用 SAS 硬盘进行双 IOM 操作的功耗约低 70~80W。原因在于，与使用 SAS 接口进行单 IOM 操作相比，单个 SATA 硬盘本身的功耗（视负载而定）大约低 0.4~0.8W。

系统考虑因素

100Gbit/秒的网络带宽和 12.5GB/秒的存储带宽，匹配效果良好。如果需要系统具备高顺序性能，我们建议设置双 IOM，并且推荐使用东芝 MG 系列的近线 SAS 企业级容量硬盘。

如果网络带宽不超过 25Gbit/秒，并且最大容量是主要目标，也可以配置单 IOM，并且搭配近线 SATA 硬盘，因为存储带宽通常不超过 4Gb/秒，可以匹配 25Gbit/秒的网络连接速度。

结语

PROMISE VTrak J5960 是一款易于维护的节能型 60 盘位顶部装载式 JBOD。比同类产品相比，它的结构非常紧凑，总长度仅为 666 毫米。在完整配备东芝 18TB 硬盘的情况下，总容量高达 1 PB，而功耗只有 500W 左右。

在这款 JBOD 上进行的存储配置评估中，东芝硬盘展现了出色的聚合性能，60 个硬盘可以提供高达 15GB/秒的顺序吞吐量和超过 30k 的随机 IOPS。JBOD 的高效冷却和气流管理技术能够确保硬盘在完全运行时，始终保持硬盘温度比环境温度低 14°C，从而确保硬盘较长的生命周期和高可靠性。

致合作伙伴的感谢信

这份实验室报告能够取得成功，离不开各位合作伙伴的大力支持和通力协作。“我想感谢所有合作伙伴对本项目的支持。PROMISE 为我们提供了这款环保 JBOD Vtrak J5960，Microchip 提供了 RAID 控制器 Adaptec SmartRAID Ultra 3254-16e/e，Broadcom 提供了 Host-Bus-Adapter HBA 9500-16e。借助所有这些优秀的产品，再加上东芝硬盘，我才得以在实验室中构建一个数据中心设置，从而让性能更上一层楼。”

Rainer Kaese, Senior Manager Business Development, Storage Products Division, Toshiba Electronics Europe GmbH

TOSHIBA

Toshiba Electronic Devices & Storage Corporation
东芝电子元件及存储装置株式会社

https://toshiba.semicon-storage.com/cn/storage.html