TOSHIBA

實驗報告

PROMISE VTrak J5960 JBOD

簡介

對於現今的數據中心而言，致力實現氣候中和及永續性，是至關重要的議題。他們的目標是在全天24小時不間斷處理和存取大量數據的同時，能維持低耗電量並降低整體碳足跡。

從廢熱回收及自然冷卻等技術，到採用於製造過程中貫徹永續理念且具備最先進技術的硬碟 (HDD)產品，環保意識興起已經對數據中心產生全面性的影響。

以下是本實驗報告所探討的產品：

PROMISE Technology 的新型 VTrak J5960 – 為一部4U 60槽的JBOD儲存擴充櫃。VTrak J5960以「具備綠色 DNA 的 JBOD」為宣傳特色，承諾落實環境保護和永續生產。東芝電子歐洲公司Toshiba Electronics Europe GmbH (以下稱「Toshiba」) 有機會於Toshiba德國實驗室中，測試和評估這款JBOD，並將其配裝上60台18TB的 Toshiba 企業級硬碟，總容量達1080TB。

Toshiba 針對功能、效能、噪音和耗電量進行測試，並側重評估 PROMISE 所聲稱的綠色設計特色。

圖片 1： Toshiba 德國實驗室中的PROMISE VTrak J5960 JBOD。

尺寸及機械特性：

J5960 的高度符合4U伺服器機架，機箱長度僅 666 公釐，是我們在 Toshiba德國實驗室中見過長度最短且高密度的 JBOD。這款 JBOD 與典型的 66 公分 2U 伺服器機箱一樣長，可輕鬆安裝進任何現有機架。與其他 JBOD 相比，這是一項很大的優勢。許多 JBOD 的長度超過 1000 公釐，需要更長的機架，也容易產生佈線問題。

J5960 將熱插拔 IO 模組 (IOM) 移到 JBOD 正面，而連接的纜線則維持在裝置的背面。如此一來，IOM 的現場更換變得極其簡單，相較之下，典型的後置替換需要穿過一堆電源和訊號線。

JBOD 的上蓋設計是有趣的機械特性之一。這款上蓋與機架連接，並在 JBOD 取出進行維修時停留在機架上。我們已對此進行測試，整體運作非常順暢。由於無需抬起上蓋，這款 JBOD 只需要拉到故障硬碟所需的位置，也可以安裝在機架的上方。

每部硬碟都使用 4 個螺絲固定至金屬支架。

LED狀態指示燈通常為關閉的，只有在上蓋取下時 (JBOD 從機架取出時) 才會開啟，因此能節省好幾瓦的額外耗電量。

Toshiba 實驗室的操作設定

機型：	PROMISE VTrak J5960 4U-SAS-60-D BP
韌體：	1023
主機作業系統：	Linux (Centos 7.9)
主機作業系統：	Windows (Windows Server 2019 Standard)
主機匯流排配接卡 (HBA)：	Broadcom Avago HBA 9500-16e (Host IF: 8x PCIe-Gen4)
RAID 配接卡：	Microchip Adaptec® SmartRAID Ultra 3254-16e/e (16x PCIe-Gen4)

測試中使用的企業級容量型硬碟：

產品型號:	Toshiba MG09SCA18TE
區塊大小:	512B emulated
韌體:	0104

圖片2: J5960內部配置.

資料速率: 282MB/s

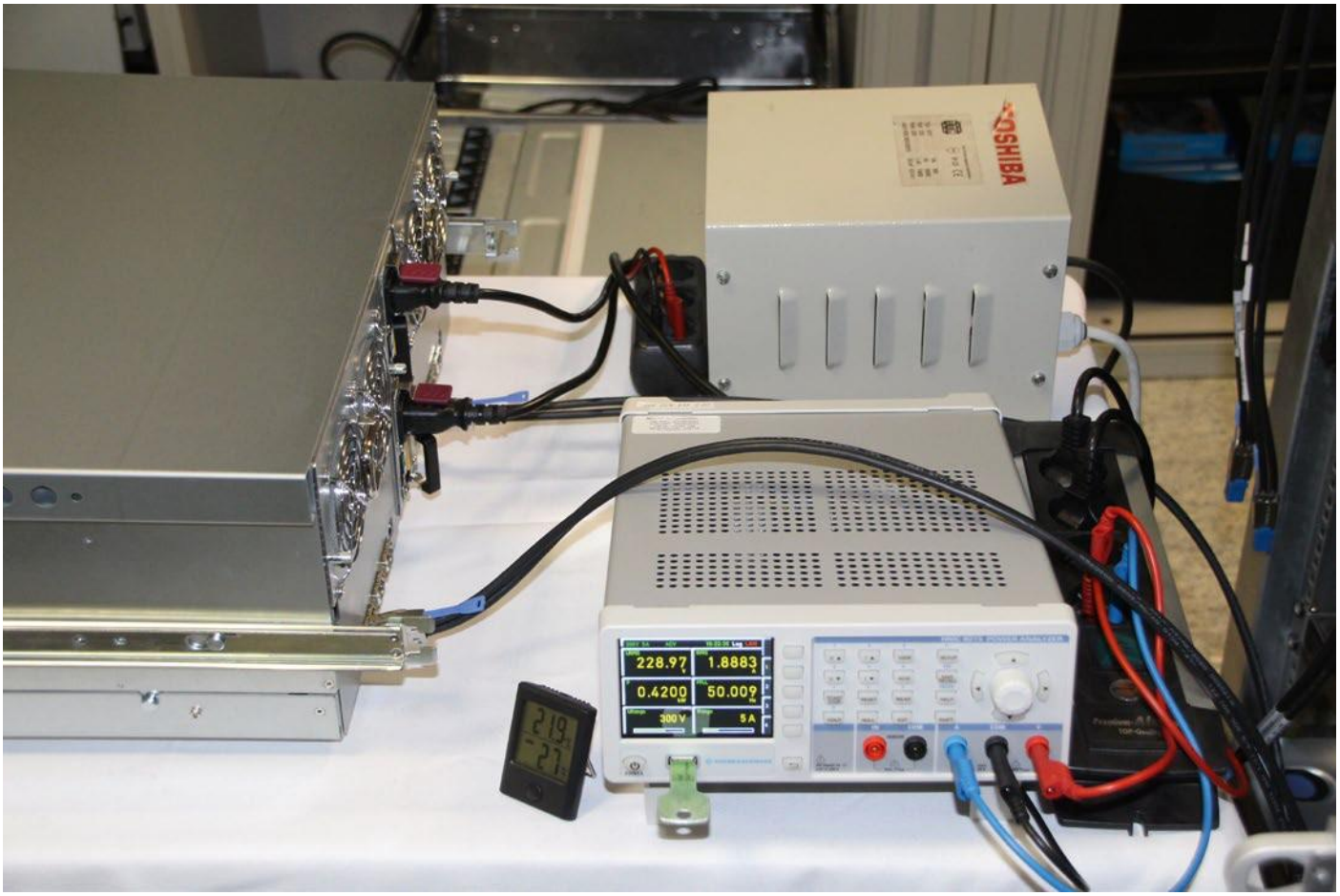
耗電量

Idle_B:	3.36W
循序寫入:	7.62W
循序讀取:	8.71W
隨機寫入:	6.64W
隨機讀取:	9.47W

JBOD 基本功能:

基本功能:	ok
SAS IOM 偵測:	ok
熱插拔/重新插入:	ok
智慧讀取:	ok
機箱管理:	ok, 使用 RJ11 序列連線測試

圖片 3：安裝於JBOD中的Toshiba MG09SCA18TE硬碟。



圖片 4：Toshiba 德國實驗室的功率測量設定。

TOSHIBA

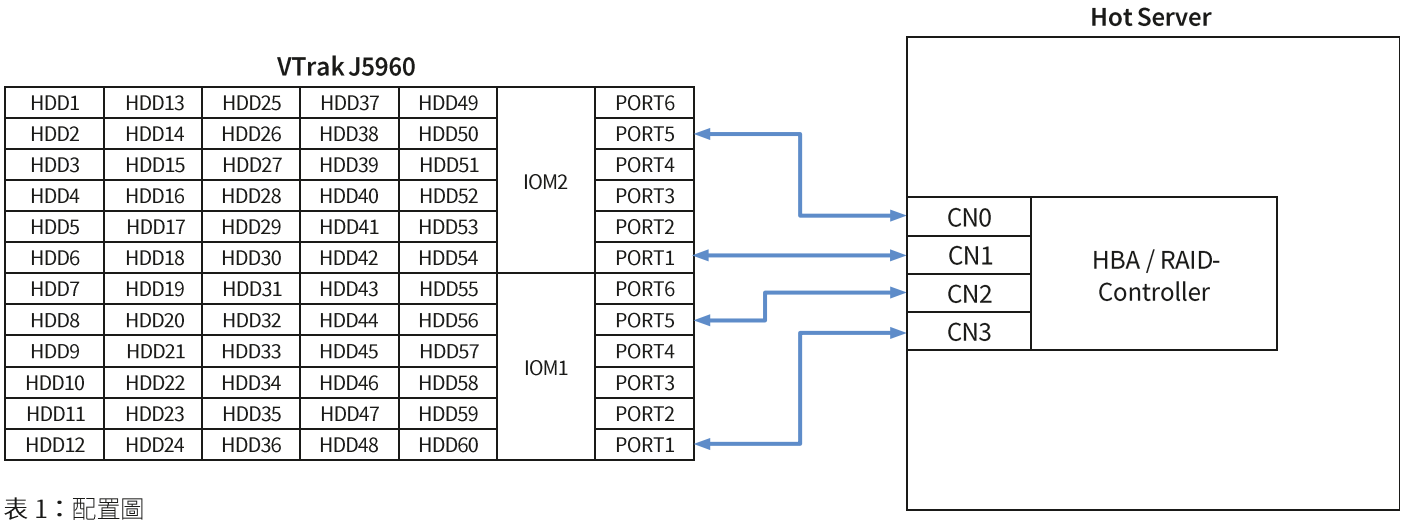


表 1：配置圖

為精準測量耗電量，我們使用高精度專業功率分析儀 (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 有一項預設的設定，若將近 2 分鐘沒有任何存取操作，所有 (SAS) 硬碟都會進入閒置狀態。對於不含硬碟的雙 IOM JBOD，100W 左右是極好的功率值。在相同條件下，專用的單 IOM JBOD 可低達 80W，但雙 IOM JBOD 一般介於 200-300W 的範圍內。

大約 300W 的「含硬碟的idle值」也是非常低的，其他 60 槽 JBOD 的起始功率為 400W 以上。0.96 的高 Lambda 係數意味著供電裝置產生的無效功率比率極低 (相當於總功率的 4%)。Lambda 係數越高 (等同於越好)，無效功率越低。無效功率不會耗費亦不會產生熱量，但供電導軌的尺寸必須按照有效功率和無效功率調整，因此高 Lambda 係數對於大型數據中心是一項重要因素。

Toshiba 實驗室的效能測量

針對 SAS 硬碟的測試，我們使用兩條 mini-SAS-HD 連接線，將 JBOD 的兩個 IOM 連接到 16e HBA 和 RAID 控制器的 4 個 mini-SAS-HD 連接埠。

此配置提供的理論 JBOD/硬碟存取頻寬為 4x 4.8GB/s = 19.2GB/s，但需要透過安裝多路徑功能進行路徑聚合。至於使用 HBA 的配置，必須在 Linux/Windows 中手動啟用多路徑。RAID 控制器 (例如 Microchip Adaptec® Ultra-3254 機型) 將自動偵測配置並啟動正確的多路徑設定。手動多路徑和 SAS 鏈路聚合僅適用於 SAS 硬碟。使用 SATA 硬碟的配置仍支援 60 部硬碟的 IOPS 聚合，但循序頻寬通常限於一個 mini-SAS-HD 鏈路 (4.8GB/s)。

我們使用「fio」(彈性的 IO 測試軟體) 測試多種硬碟配置，測量了循序、隨機和混合工作負載效能及其耗電量。此外，已針對透過 HBA 連接的個別硬碟及採用 RAID 配置的實體硬碟，以及邏輯硬碟進行測試。針對邏輯硬碟，我們還測量了複製 (即讀取和寫入) 大型檔案的效能和功率。

JBOD 設定備註

VTrak J5960 預設為「硬碟 2 分鐘無活動即進入閒置狀態」。閒置硬碟從閒置狀態轉換到運作狀態，大約需要 1200 毫秒。
建議針對 RAID 配置停用此功能，因為在大型 RAID 中，即使 RAID 處於完整載入資料的狀態，部分硬碟可能在 2 分鐘後停止運作。若 RAID 中的部分硬碟設定為閒置模式，則存取這些硬碟時會出現 1200 毫秒的長延遲。
停用閒置模式切換：透過序列連接線連接至 IOM 管理連接埠 (115200/8/N/1)，CLI 指令為「enclosure -m -idlep 0」。

Toshiba 測試使用「zoning mode 0」(= 預設設定) 進行，這表示「無分區」，因此可以從兩個 IOM 存取所有硬碟。

部分 RAID 控制器或 HBA 可能拒絕將 4 條連接線連接至兩個 IOM。若遇到這樣的情況，「zoning mode 1」可能是一種解決方法。此模式下，30 部硬碟從一個 SAS 連接埠連接至 IOM1，另外 30 部硬碟則從一個 SAS 連接埠連接至 IOM2。配置看起來就像是兩個裝有 30 部硬碟的 JBOD。將分區模式 CLI 指令變更為「enclosure -m -z 1」。

所有硬碟作為個別實體裝置平行運作 (多路徑):

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

工作負載	功率(W)	IOPS	頻寬(MB/s)
循序寫入 1024K	610		13300
循序讀取 1024K	640		14500
隨機寫入 4K	510	24100	
隨機讀取 4K	540	33900	
混合型 4K/64K/256K/2M	540	22600	2350
環境溫度	23°C		
最低硬碟溫度	27°C		
最高硬碟溫度	36°C		

理論最大頻寬為 282MB/s (單一磁碟) x 60 = 16.2GB/s。憑藉 14.5GB/s 和 20K+ IOPS，此配置接近理論極限。

最大 640W 的功率為 JBOD 的綠色認證提供有力證明。最熱硬碟和最冷硬碟之間的溫差小於 10°C，最高環境溫度低於 14°C，可提供高效散熱，有助硬碟維持可靠性和延長使用壽命。

所有硬碟作為 RAID10，Windows:

作業系統: Windows Server 2019
RAID/配接卡: Microchip Adaptec® SmartRAID Ultra 3254-16e/e (16x PCIe-Gen4)
硬碟: 60x Toshiba MG09SCA18TE
配置: Dual IOM 2x 2 Mini-SAS HD cables (3m in length)

工作負載	功率(W)	IOPS	頻寬(MB/s)
循序寫入 1024K	510		8600
循序讀取 1024K	570		15300
隨機寫入 4K	600	12800	
隨機讀取 4K	730	9900	
混合型4K/64K/256K/2M	680	6100	1800
閒置 (RAID 背景運作)	470		
環境溫度	25°C		
最低硬碟溫度	29°C		
最高硬碟溫度	38°C		

指令碼 1 – 所有硬碟作為個別實體裝置平行運作 (多路徑):

```
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}
```

指令碼 2 – 所有硬碟作為 RAID10，Windows 實體硬碟:

```
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
```

由於在 RAID10 中總是平行寫入兩部鏡像裝置，因此寫入速度下降一半。RAID 配置中的閒置功率幾乎等同於運作中的功率，因為 RAID 控制器持續在背景存取硬碟進行一致性檢查。

工作負載	功率(W)	IOPS	頻寬(MB/s)
循序寫入 1024K	520		6900
循序讀取 1024K	550		15000
隨機寫入 4K	520	11100	
隨機讀取 4K	540	29500	
混合型4K/64K/256K/2M	550	8100	2400
Windows 複製	500		550
閒置 (RAID 背景運作)	470		
環境溫度	25°C		
最低硬碟溫度	29°C		
最高硬碟溫度	39°C		

所有硬碟作為 RAID10，Windows 邏輯磁碟區:

作業系統:	Windows Server 2019
RAID/配接卡:	Microchip Adaptec® SmartRAID Ultra 3254-16e/e (16x PCIe-Gen4)
硬碟:	60x Toshiba MG09SCA18TE
配置:	Dual IOM 2x 2 Mini-SAS HD cables (3m in length)

指令碼 3 – 所有硬碟作為 RAID10，Windows 邏輯磁碟區:

```
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
```

此針對 Windows 中邏輯磁碟區的基準測試並未使用整個硬碟，而是對 1TB 的測試檔案大小執行。因此，其更貼近於實際的使用案例。IOPS 超過 32k 是因為尋軌操作未涵蓋整個容量範圍。相較於前述對實體硬碟進行完整 500TB 尋軌，這也顯著降低寫入操作的耗電量。

SATA 硬碟的配置

我們還使用 SATA 硬碟 (型號 MG09ACA18TE) 測試 VTrak J5960 JBOD。由於 SATA 介面只有一個訊號路徑，我們使用單 IOM 配置 (IOM2 已取出，所有四條連接線都連接至 IOM1)。

循序效能通常僅限於一條 mini-SAS HD 連接線的水準 (循序讀寫約為 4.3GB/s)。IOPS 值與 SAS 的結果非常適配，因為它們不受頻寬限制 (全容量為 10k IOPS，1TB 資料範圍的邏輯硬碟為 30k IOPS)。

SATA 硬碟和單 IOM 操作的功耗比使用 SAS 硬碟和雙 IOM 的同等設定低約 70 至 80W。原因在於單個 SATA 硬碟的功耗比採用 SAS 介面和缺少第二個 IOM 的相同硬碟低約 0.4 至 0.8W (取決於負載)。

系統考量

100Gbit/s 的網路頻寬和 12.5GB/s 的儲存頻寬確實非常適配。對於需要如此高循序效能的系統，我們建議在雙 IOM 設定中使用 Toshiba MG 系列的近線 SAS 企業級容量型硬碟。

若網路頻寬為 25Gbit/s 以下，並且以最高容量為主要目標，也可以使用單 IOM 配置的近線 SATA 硬碟，其儲存頻寬通常限制為 4GB/s，同樣與 25Gbit/s 的網路連結速度適配。

摘要

PROMISE VTrak J5960 是一款節能且易於維護並擁有60個硬碟插槽的JBOD。為同級產品中最輕巧的機型，總長度僅 666 公釐。將所有插槽安裝Toshiba 18TB硬碟後，可在僅約 500W 的功耗下，提供超過 1 PB 的總容量。

Toshiba 已針對採用此 JBOD 的儲存配置進行評估，得出 60 部硬碟的聚合效能 (最高 15GB/s 連續處理量和超過 30k 隨機 IOPS)，而 JBOD 的高效散熱和氣流管理可協助硬碟延長使用壽命和維持高可靠性，在完全運轉時溫度始能終低於環境溫度 14°C。

感謝我們的合作夥伴

我們的合作夥伴是本實驗報告得以成功的關鍵。「我要感謝所有合作夥伴對此專案的協助：PROMISE 提供經綠色認證的 JBOD Vtrak J5960，Microchip 提供 RAID 控制器 Adaptec SmartRAID Ultra 3254-16e/e，Broadcom 提供主機匯流排配接卡 HBA 9500-16e。最後加上我們的 Toshiba 硬碟，我們在實驗室成功設置資料中心環境，展現全新境界的傑出效能成果。」

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

TOSHIBA

Toshiba Electronic Components Taiwan Corporation
台灣東芝電子零組件股份有限公司

4F., No.168, Sec.3, Nan-jing E. Rd., Taipei, 104105 Taiwan
104105 台灣台北市中山區南京東路三段168號4樓

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