FUJITSU

FUJITSU Software
Technical Computing Suite V4.0L20

Job Operation Software Administrator's Guide for HPC Extensions

J2UL-2460-02ENZ0(01)
November 2021

Preface

Purpose of This Manual

This manual is an administrator's guide to the necessary functions for FX server operation. These functions are part of the HPC (High Performance Computing) extension function included in the Job Operation Software of Technical Computing Suite.

The manual describes a tool for collecting statistical information for large pages and efforts to improve the execution performance of parallel applications.

Intended Readers

This manual is intended for administrators who operate and manage the FX server.

The manual assumes readers have the following knowledge:

- Basic Linux knowledge
- Overall knowledge of the Job Operation Software, obtained from the "Job Operation Software Overview"

Organization of This Manual

This manual is organized as follows.

"Chapter 1 Overview of the HPC Extension Function"

This chapter describes the listed functions provided by the HPC extension function.

"Chapter 2 Large Page Statistical Information Collection Tool (lpgstat)"

This chapter describes the necessary system settings for enabling the Fujitsu HPC extension function for the FX server.

"Appendix A Efforts to Improve the Execution Performance of Parallel Applications"

This appendix describes the efforts made to improve the execution performance of parallel applications on the FX server.

Notation in This Manual

Representation of units

The following table lists the prefixes used to represent units in this manual. Basically, disk size is represented as a power of 10, and memory size is represented as a power of 2. Be careful about specifying them when displaying or entering commands.

Prefix	Value	Prefix	Value
K (kilo)	10^3	Ki (kibi)	2^{10}
M (mega)	10^6	Mi (mebi)	2^{20}
G (giga)	10^9	Gi (gibi)	2^{30}
T (tera)	10^{12}	Ti (tebi)	2^{40}
P (peta)	10^{15}	Pi (pebi)	2^{50}

Representation of Model Names

In this manual, a computer with a mounted Fujitsu CPU A64FX is abbreviated as "FX server".

Symbols in This Manual

This manual uses the following symbols.

Example

The Example symbol indicates an application example.

See

The See symbol indicates the reference source of detailed information.

Export Controls

To export or release this document to a third party, check and take the necessary procedures in accordance with the applicable laws and regulations of your resident country and U.S. export control laws.

Trademarks

Linux(R) is a registered trademark of Linus Torvalds in the U.S. and other countries.

Other company names and product names appearing in this manual are trademarks or registered trademarks of their respective owners.

Date of Publication and Version

Version	Manual Code
November 2021, Version 2.1	J2UL-2460-02ENZ0(01)
March 2020, 2nd version	J2UL-2460-02ENZ0(00)
January 2020, First version	J2UL-2460-01ENZ0(00)

Copyright

Update history

Changes	Location	Version
Changed the description of kernel boot parameters.	A.1, A.2	2.1
Changed the description of system setting items.	A.1	
Changed the look according to product upgrades.	-	2

Contents

Chapter 1 Overview of the HPC Extension Function

This chapter describes the function (HPC extension function) supporting the use of FX server-specific functions in the OS and Technical Computing Suite.

1.1 HPC Extension Function List

This section lists the administrator functions provided by the HPC extension function.

- Large page library
  The library provides a tool for administrators to collect statistical information for large pages.
- Various drivers/libraries
  The various provided drivers support the FX server.
  - TofuD driver
  - Power control driver/library
  - Inter-core hardware barrier driver/library
  - Sector cache driver/library
- Dump generation management
  A utility is provided so that compute node resources can be used efficiently by appropriately controlling the number of maintenance data sets (memory dumps) collected by the FX server.

In addition to the above, efforts have been made to improve the execution performance of parallel applications for the FX server.

Out of the above, this manual describes the tool for administrators to collect statistical information for large pages and the efforts to improve the execution performance of parallel applications.

See

The HPC extension function operates in linkage with individual functions provided by the Job Operation Software of Technical Computing Suite. For the positioning of the HPC extension function in the Job Operation Software of Technical Computing Suite, see the "Job Operation Software Overview" manual.

Chapter 2 Large Page Statistical Information Collection Tool (lpgstat)

This chapter describes the tool for collecting statistical information for large pages (lpgstat). The tool is part of the HPC extension function.

2.1 lpgstat Command

The lpgstat command is provided together with the large page library of the FX server. When executed on the FX server, this command outputs the use status of system memory and job memory for all nodes or each NUMA node. Also, the FX server has a setting for using the lpgstat command to regularly collect large page statistical information.

Command Specifications

NAME

lpgstat - Command that displays the use status of system memory and job memory for all nodes or each NUMA node

SYNOPSIS

/opt/FJSVxos/sbin/lpgstat [{-n|-v}]

OPTIONS

-n
This option displays the use status of system memory and job memory for all nodes, at the standard output.

-v
This option displays the use status of system memory and job memory for each NUMA node and information on mounted CPUs, at the standard output.
If no option is specified, the command operates with the default option (-n).

Table 2.1 Displayed Contents of the lpgstat Command

Displayed Contents	Description
NUMA Node=<node-id>	Node number displayed when data is displayed for each NUMA node (when the -v option is specified)
CPU List=<cpulist>	Online CPU numbers displayed when data is displayed for each NUMA node (when the -v option is specified)
[System]	Type of information on memory use (system memory)
System_MemTotal	Total amount of system memory (KiB)
System_MemFree	Amount of free system memory (KiB)
System_MemUsed	Amount of system memory in use (KiB)
[Job]	Type of information on memory use (job memory)
Job_MemTotal	Total amount of job memory (KiB)
Job_MemFree	Amount of free job memory (KiB)
Job_MemUsed	Amount of job memory in use (KiB)
Job_MemUsed_normal	Amount of job memory in use (KiB) as normal pages
Job_MemUsed_2MB	Amount of job memory in use (KiB) as large pages (2-MiB pages)

END CODE

The command returns the following end codes.
0: Normal end
Other than 0: Abnormal end

Example

Examples of lpgstat output are shown below.

Normally, the lpgstat command is used to regularly collect statistical information, but the administrator may be directly executing commands on the FX server. If so, the administrator is requested to use the batch execution function (pmexe command) to collect output.

- Output of the amount of memory used by all nodes (when "lpgstat -n" is executed)

```
[System]
    System_MemTotal:                        8,800,064 KB
    System_MemFree:                         8,023,040 KB
    System_MemUsed:                           777,024 KB
[Job]
    Job_MemTotal:                          23,294,592 KB
    Job_MemFree:                           21,885,568 KB
    Job_MemUsed:                            1,409,024 KB
    Job_MemUsed_normal:                     1,404,928 KB
    Job_MemUsed_2MB:                            4,096 KB
```

- Output of the amount of memory used by each NUMA node (when "lpgstat -v" is executed)

```
[System]
NUMA Node=0                   CPUList=0
    System_MemTotal:                        1,147,456 KB
    System_MemFree:                           807,872 KB
    System_MemUsed:                           339,584 KB
NUMA Node=1                   CPUList=1
    System_MemTotal:                        2,552,576 KB
    System_MemFree:                         2,154,688 KB
    System_MemUsed:                           397,888 KB
- Omitted -
[Job]
NUMA Node=4                   CPUList=12,13,14,15,16,17,18,19,20,21,22,23
    Job_MemTotal:                           5,825,152 KB
    Job_MemFree:                            4,921,152 KB
    Job_MemUsed:                              904,000 KB
    Job_MemUsed_normal:                       764,736 KB
    Job_MemUsed_2MB:                          139,264 KB
NUMA Node=5                   CPUList=24,25,26,27,28,29,30,31,32,33,34,35
    Job_MemTotal:                           5,825,216 KB
    Job_MemFree:                            5,549,312 KB
    Job_MemUsed:                              275,904 KB
    Job_MemUsed_normal:                       275,904 KB
    Job_MemUsed_2MB:                                0 KB
- Omitted -
```

See

For details on the batch execution function (pmexe command), see "Overview of the Operation Support Function" in the "Job Operation Software System Management" manual.

Regular Collection of Large Page Statistical Information

On the FX server, cron is used to set the regular collection of large page statistical information.

The execution of "lpgstat -n" once every 10 minutes outputs large page statistical information to the following file.

```
/var/log/FJSVxos/mmm/lpgstat.log
```

The file is rotated at 2 MB. When the number of rotations is 10, the oldest file is deleted.

Example

The following example shows output to lpgstat.log.

```
Tue Apr 16 15:20:02 JST 2019
[System]
    System_MemTotal:                          8,800,064 KB
    System_MemFree:                           8,023,040 KB
    System_MemUsed:                             777,024 KB
[Job]
    Job_MemTotal:                            23,294,592 KB
    Job_MemFree:                             21,885,568 KB
    Job_MemUsed:                              1,409,024 KB
    Job_MemUsed_normal:                       1,404,928 KB
    Job_MemUsed_2MB:                              4,096 KB
Tue Apr 16 15:30:01 JST 2019
[System]
    System_MemTotal:                          8,800,064 KB
    System_MemFree:                           8,023,040 KB
    System_MemUsed:                             777,024 KB
[Job]
    Job_MemTotal:                            23,294,592 KB
    Job_MemFree:                             21,885,568 KB
    Job_MemUsed:                              1,409,024 KB
    Job_MemUsed_normal:                       1,404,928 KB
    Job_MemUsed_2MB:                              4,096 KB
Tue Apr 16 15:40:02 JST 2019
- Omitted -
```

Appendix A Efforts to Improve the Execution Performance of Parallel Applications

This appendix describes the efforts made to improve the execution performance of parallel applications on the FX server.

The processing time of synchronously executed parallel applications tends to lengthen due to adverse effects on performance from the system noise of daemon operations, hardware interrupts, etc. There is also a possibility of a variation in performance occurring due to timing, for example, under cache conditions when executing certain processing.

Therefore, to reduce system noise and suppress performance variations on the FX server, the measures described in the appendix are taken at the time of system build.

The appendix also describes boot parameter setting items changed from those in the standard Linux kernel to support the FX server configuration.

Avoid changing the settings described in the appendix since it may affect job operation and execution performance.

A.1 Measures to Reduce System Noise and Suppress Performance Variations

This section describes the set items for reducing system noise and suppressing performance variations.

Boot Parameter Settings of the Kernel

The following table lists the boot parameter settings for the kernel to reduce system noise and suppress performance variations.

Table A.1 Boot Parameters of the Kernel

Boot Parameter	Meaning	Effect	Remarks
nohz_full=12-59	Enables the dynamic tickless function on the specified CPUs. When the CPU run queue has only 1 task, the dynamic tickless function changes the clock tick interval from 10 milliseconds to 1 second.	Reduces system noise	Specifies CPUs (12 - 59) that are set for job processes.
nosoftlockup	Disables detection of soft-lockup, which is executed regularly.	Reduces system noise	
audit=0	Disables the audit subsystem and audit log collection.	Reduces system noise	
mce=ignore_ce	Disable features for corrected errors, e.g. polling timer and CMCI(Corrected Machine Check Interrupt).	Reduces system noise	
norandmaps	Disables randomization of virtual address spaces for applications.	Suppresses performance variations	

Daemon Process Stopped

To reduce system noise, the FX server stops the following daemon process.

Table A.2 Stopped Daemon Process

Daemon Name	Effect of Stopping
irqbalance	Dynamic allocation processing by irqbalance is disabled since the FX server uses assistant cores to accept hardware interrupts.

System Settings at Startup

To reduce system noise, the FX server uses the following system settings at node startup.

Table A.3 System Setting Items

System Setting Item	Setting Description
Worker thread CPU mask	Write 0xf in the files listed below to set the CPU mask. Of the asynchronous processing queues of the kernel (workqueue), those that do not need to operate in specific cores are bound to the system cores (0 - 3). /sys/devices/virtual/workqueue/cpumask /sys/devices/virtual/workqueue/writeback/cpumask

A.2 FX Server-Specific Boot Parameters of the Kernel

The following table describes boot parameters changed from or added to those in the standard Linux kernel to support the FX server configuration.

Table A.4 FX Server-Specific Boot Parameters

Boot Parameter	Meaning	Remarks
crashkernel=128M	Sets the crashkernel area with a size of 128 MiB.	Set for maintenance purposes.
crash_kexec_post_notifiers=1	Sets execution at the panic time, starting with panic-notifiers and then crash_kexec processing.	Set for maintenance purposes.
earlycon=pl011,0x1c050000 console=ttyAMA0	Enables log output to the console device from the initial stage of starting the kernel.	Set for maintenance purposes.
tofroot=<*BIO IP address*>:<*root file path of each BIO compute node*>	Added parameters for the FX server. They set the paths to the root files required for starting compute nodes (diskless nodes).	
hugepagesz=2M hugepagesz=512M default_hugepagesz=2M	Sets the memory page sizes for using large pages.	
acpi=force	The ACPI table is used in preference to the device tree when booting the kernel.	
transparent_hugepage=never	Disable the Transparent Huge Page.	
ipmi_msghandler.panic_op=none	Stop logging to the BMC when the kernel panic occurs.	
arm_smmu_v3.disable_bypass=0	Set to bypass the IO virtual memory translation.	