NEC

NEC Network Queuing System (NQSV)
移行ガイド

輸出する際の注意事項

本製品（ソフトウェアを含む）は、外国為替および外国貿易法で規定される規制貨物（または役務）に該当することがあります。

その場合、日本国外へ輸出する場合には日本国政府の輸出許可が必要です。

なお、輸出許可申請手続きにあたり資料等が必要な場合には、お買い上げの販売店またはお近くの当社営業拠点にご相談ください。

は　し　が　き

本書は、NQSII R3.00 / NQSII R4.00 から NQSV への移行方法および、NQSV の R1.02 までの主な新規機能、差分について説明したものです。

本書は、NQSV を新規に導入後、既存の NQSII システムを NQSV に移行することを前提としています。

備考

(1)　本書はNEC Network Queuing System V (NQSV) R1.00以降に対応しています。

(2)　本書に説明しているすべての機能はプログラムプロダクトであり、以下のプロダクト名およびプロダクト番号に対応しています。

プロダクト名	型番
NEC Network Queuing System V (NQSV) /ResourceManager	UWAF00 UWHAF00（サポートパック）
NEC Network Queuing System V (NQSV) /JobServer	UWAG00 UWHAG00（サポートパック）
NEC Network Queuing System V (NQSV) /JobManipulator	UWAH00 UWHAH00（サポートパック）

(3)　UNIX は The Open Group の登録商標です。

(4)　OpenStackは、アメリカ合衆国およびその他の国における OpenStack Foundation の商標です。

(5)　Red Hat OpenStack Platformは、アメリカ合衆国およびその他の国における Red Hat, Inc. の商標です。

(6)　Linux は Linux Torvalds氏の米国およびその他の国における登録商標あるいは商標です。

(7)　Dockerはアメリカ合衆国およびその他の国におけるDocker, Inc.の商標です。

(8)　InfiniBand は、InfiniBand Trade Associationの商標またはサービスマークです。

(9)　その他、記載されている会社名、製品名は、各社の登録商標または商標です。

本書の読み進め方

本書は、次の構成となっています。章ごとに対象読者の範囲は異なっており、表の一番右の列にその範囲を示しています。記載された対象読者の後に(*)がついている章については、該当する読者は必ずお読みください。

–

章	タイトル	内容	対象読者
1	移行手順	NQSII R4.00/R3.00 から NQSV への移行手順についての説明	システム管理者(*)
2	NQSII R4.00 からの差分	NQSII R4.00 と NQSV の差分についての説明	システム管理者(*)
3	NQSII R3.00 からの差分	NQSII R3.00 と NQSII R4.00 の差分についての説明	システム管理者(*)

関連説明書

NEC Network Queuing System V (NQSV)のマニュアルは以下で構成されています。

マニュアル名称	内容
NEC Network Queuing System V (NQSV) 利用の手引［導入編］	システムの全体像および基本的なシステムの構築方法に関する説明
NEC Network Queuing System V (NQSV) 利用の手引［管理編］	管理者が実施する各種設定に関する説明
NEC Network Queuing System V (NQSV) 利用の手引［操作編］	一般利用者が使用する各種機能に関する説明
NEC Network Queuing System V (NQSV) 利用の手引［リファレンス編］	コマンドリファレンス
NEC Network Queuing System V (NQSV) 利用の手引［API 編］	NQSV を操作するプログラミングインタフェース（API）に関する説明
NEC Network Queuing System V (NQSV) 利用の手引［JobManipulator 編］	スケジューラコンポーネント JobManipulator に関する説明
NEC Network Queuing System V (NQSV) 利用の手引［アカウンティング・予算管理編］	アカウンティング機能に関する説明

表記上の約束

本書では次の表記規則を使用しています。

省略記号 ...　　前述の項目を繰り返すことができることを表しています。ユーザは同様の項目を任意の数だけ入力することができます。

縦棒　　|　　オプションまたは必須の選択項目を分割します。

中かっこ {}　　１つを選択しなければならない一連パラメータまたはキーワードを表しています。

角かっこ [ ]　　省略可能な一連パラメータまたはキーワードを表しています。

用語定義・略語

用語・略語	説　明
ベクトルエンジン (VE、Vector Engine)	SX-Aurora TSUBASAの中核であり、ベクトル演算を行う部分です。PCI Expressカードであり、x86サーバーに搭載して使用します。
ベクトルホスト (VH、Vector Host)	ベクトルエンジンを保持するサーバー、つまり、ホストコンピュータを指します。
IB	InfiniBandの略語です。
HCA	Host Channel Adapterの略語です。IBネットワークに接続するためにサーバー側に取り付けるPCIeカードです。
MPI	Message Passing Interfaceの略語です。主にノード間で並列コンピューティングを行うための標準化規格です。

目　次

第1章　　移行手順

1.1　NQSII R4.00 からの移行

1.1.1　移行手順

NQSII R4.00 からの NQSV への移行手順は、以下の通りです。

(1)　NQSVのBSV環境を準備する

(2)　実行ホストの NQSII/JobServer をアンバインドして停止します。

(3)　実行ホストの NQSII/JobServer をアンインストールします。

(4)　実行ホストに NQSV/JobServer をインストールします。

(5)　NQSV/JobServer を起動し、NQSV/BSVにバインドします。

1.1.2　注意事項

アカウントデータ

- NQSIIのアカウントデータをNQSVに移行することはできません。
- 移行前の退避したアカウントデータは、NQSII R4.00のアカウントコマンドで参照してください。
  退避方法は以下の通りです。

  NQSII 利用の手引[アカウンティング・予算管理編] 3.5 アカウントデータの退避
- 移行後のアカウントデータはNQSVで管理します。

1.2　NQSII R3.00 からの移行

1.2.1　移行手順

NQSII R3.00 からの NQSV への移行手順は、以下の通りです。

(1)　NQSVのBSV環境を準備します

(2)　実行ホストの NQSII/JSV R3.00 をアンバインドして停止します

(3)　実行ホストの NQSII/JSV R3.00 をアンインストールします

(4)　計算ノードのOSアップデート(RHEL6→7)します

(5)　実行ホストにNQSV/JobServerをインストールします。

(6)　NQSV/JobServerをNQSV/BSVにバインドします

1.2.2　注意事項

アカウントデータ

- アカウントデータの移行することはできません。
- 移行前のアカウントデータはバックアップしてSCACCTで参照してください。退避方法は以下の通りです。

  SCACCT利用の手引 3.5 アカウントデータの退避
- 移行後のアカウントデータはNQSVで管理します。

第2章　NQSII R4.00 からの差分

2.1　新規機能

2.1.1　論理ジョブを論理ホストに変更

NQSII の「論理ジョブ」は、NQSV では「論理ホスト」となります。

(1)　リクエスト、ジョブ、論理ホストの関係

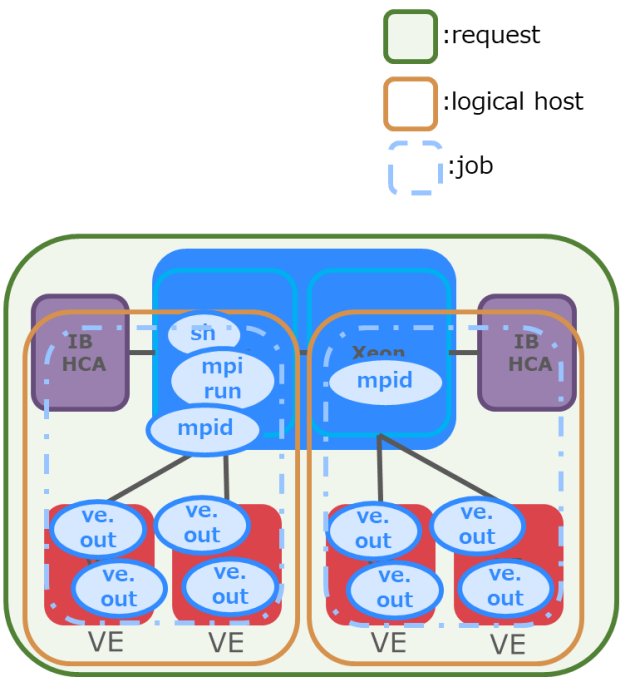
リクエストは、NQSVが管理するユーザジョブのことです。

ジョブは、実行ホスト上での実行単位で、ジョブサーバ上で実行するプロセスの集合です。

論理ホストは、実行ホストのリソースを区切って形成した仮想的なホストです。１つの実行ホストから複数の論理ホストを作ることが可能です。

```
#PBS -b 2
#PBS -T necmpi
#PBS -l elapstim_req=300
#PBS --cpunum-lhost=2
#PBS --memsz-lhost=10gb
#PBS --venum-lhost=2
#PBS -use-hca=2

mpirun -nn 2 -np 8 ./ve.out
```



(2)　機能

資源制限のオプションを「logical job」から「logical host」に変更しました。

NQSII	NQSV	
-l cpunum_job	--cpunum-lhost	論理ホストあたりのCPU台数
-l cputim_job	--cputim-lhost	論理ホストあたりのCPU使用時間制限値

-l gpunum_job	--gpunum-lhost	論理ホストあたりのGPU数
-l memsz_job	--memsz-lhost	論理ホストあたりのメモリサイズ制限
-l vmemsz_job	--vmemsz-lhost	論理ホストあたりの仮想メモリサイズ制限

NQSIIのオプションもNQSVで引き続き利用可能です。

(3)　関連説明書

NQSV 利用の手引［管理編］4. キューの管理

NQSV 利用の手引［操作編］1.2.9. 資源制限値の指定

2.1.2　SX-Aurora TSUBASA アーキテクチャサポート VE/HCA のリソース管理

(1)　qsub(1) のオプション追加

投入時に VE数、HCA数を要求できるようオプションを追加しました。以下のオプションが追加されています。

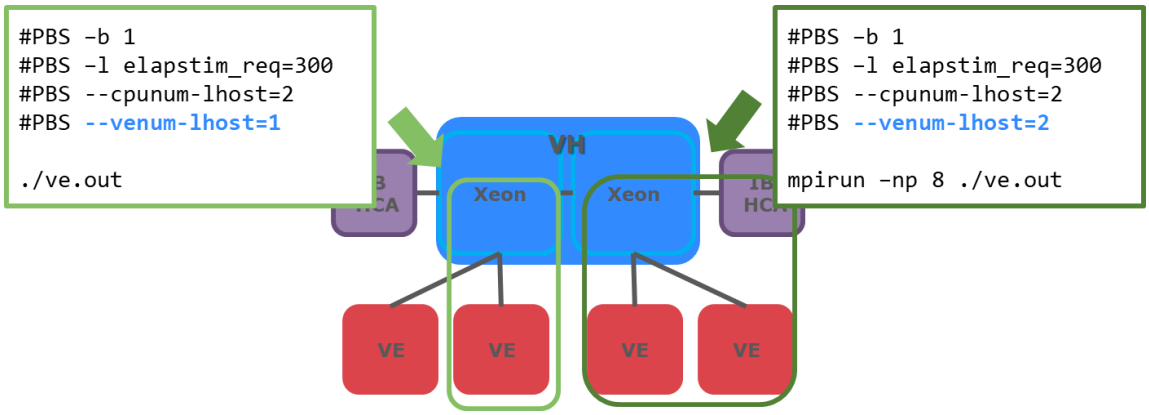
NQSV	単位	説明
--venum-lhost	論理ホスト	論理ホストあたりの VE ノード数
--venode	リクエスト	リクエストが必要とする VE ノードの総数
--use-hca	VE	同一デバイスグループの VE が使用する、HCA のポート数 VE を使用するリクエストに対してのみ有効です

これらのオプションは、実行ホストがSX-Aurora TSUBASA システムの場合のみ有効です。

--venum-lhost

--venum-lhost は論理ホストあたりのVE数を指定するオプションです。VEは、VE単位でアサインしますので、ジョブから占有されます。よって、VEノードあたりのコア数や、メモリ量を指定する必要はありません。

これらの指定方法はGPUと同じと考えてください。



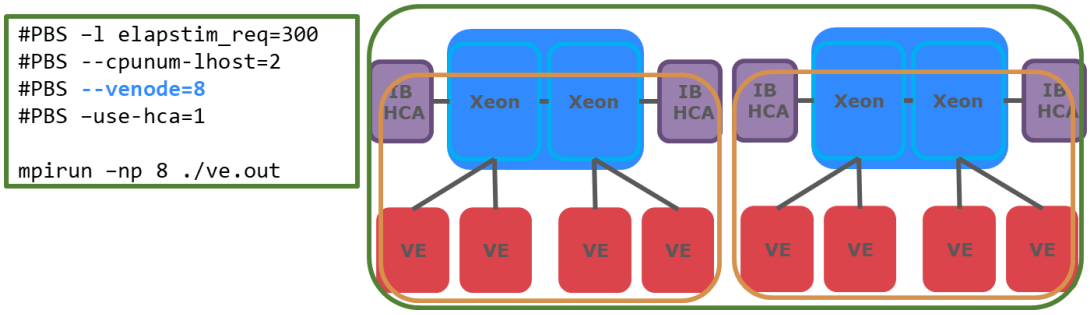
--venode

--venode は、リクエスト全体のVE数を指定するオプションです。

このオプションを指定すると、キューに設定された既定搭載VE ノード数に従って、ジョブ数が自動換算されてリクエストに適用されます。

(詳細は、NQSV 利用者の手引 [管理編]
13.1 VEの総数指定投入とキューの既定VE搭載数の設定 を参照してください)

そのため、本オプションとジョブ数の指定（-b オプション）は同時に行うこと
はできません。



--use-hca

--use-hca は、Direct通信の利用、および論理ホスト内における割り当てVE が所属するデバイスグループあたりに必要なHCAポート数を指定するオプションです。

ex) --use-hca=all:1

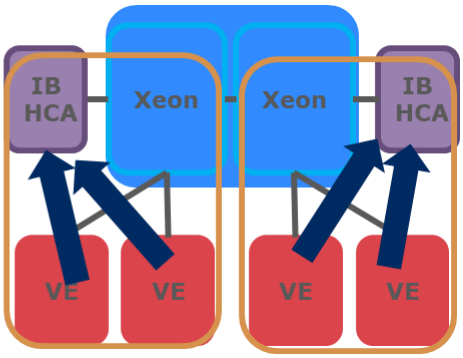
NEC MPIと連携し、通信コストが最小となる最適なHCAポートをアサインします。この機能を使用する場合、リソース定義ファイルが必要となります。

(詳細については、NQSV 利用の手引 [JobManipulator編] 5.4.2. (2) デバイスリソース定義ファイル を参照してください)

VEノードを使用し、ジョブ数が２以上の場合、このオプションに１以上の値を設定してください。

```
#PBS -b 2
#PBS -T necmpi
#PBS --cpunum-lhost=2
#PBS --use-hca=1
#PBS --venum-lhost=2

mpirun -nn 2 -np 8 ./ve.out
```



(2)　qstat(1) によるVE情報表示

qstat に -Je オプションを追加しました。

```
$ qstat -Je
JNO  RequestID        EJID  VEMemory    VECPU  JSVNO VectorIsland    UserName Exit
---- ---------------- ----- -------- -------- ----- --------------- -------- ----
   0 72.host.example 11366    1.00GB     2.00     10 exechost.exampl user1       -
```

qstat に --venode オプションを追加しました。

```
$ qstat --venode
VectorIsland VE_No  Cores Memory Status        OS_Status
------------ ----- ------ ------ ------------- ----------
host8_001        0      8    48GB ONLINE        ONLINE
host8_002        1      8    48GB ONLINE        ONLINE
host8_003        2      8    48GB ONLINE        ONLINE
host8_004        3      8    48GB ONLINE        ONLINE
host8_005        4      8    48GB ONLINE        ONLINE
host8_006        5      8    48GB ONLINE        ONLINE
host8_007        6      8    48GB ONLINE        ONLINE
host8_008        7      8    48GB ONLINE        ONLINE
```

(3)　関連説明書

NQSV 利用の手引 [管理編] 13. VE およびGPU 対応

NQSV 利用の手引 [操作編] 1.2. バッチリクエストの投入

NQSV 利用の手引 [操作編] 5.3 詳細情報の確認

NQSV 利用の手引 [JobManipulator] 5.4 HCA割り当て機能

(4)　注意事項

HCAを正しく割り当てるためには、各実行ホストにデバイスリソース定義ファイル(/etc/opt/nec/nqsv/resource.def)の設定をする必要があります。

VE/HCA 障害処理

(1)　機能

NQSVは、VE障害を検出すると、そのVEを除外し、残ったVEで運用を継続します。また、実行ホストにVE とHCA が搭載されている場合、HCA の障害を検知した際にそのJSV を自動的に運用から除外することが可能です。

(2)　関連説明書

NQSV 利用者の手引［管理編］13.4. HCA障害の検知

NEC MPI サポート

(1)　機能

NEC MPIをサポートしました。

(2)　関連説明書

NQSV 利用の手引［操作編］1.14.1. NEC MPI 環境での実行

SX-Aurora TSUBASA トポロジーを意識したスケジューリング

(1)　機能

通信コストが最小となるようなVE割り当て、HCA割り当てが可能です。

(2)　関連説明書

NQSV 利用の手引［JobManipulator編］5.4. HCA割り当て機能

VE アカウンティング 機能

(1)　機能

scacctreq(1) およびscacctjob(1)の -V オプションで、VEのアカウンティング情報が参照可能です。

(2)　関連説明書

NQSV 利用者の手引［アカウンティング・予算管理編］

2.1.3　sstat(1)　強化

(1)　機能

実行ホストの使用可能リソース制限値、省電力状態、HW障害情報をまとめて一括表示する

詳細情報表示機能を追加しました。

```
$ sstat -E –f
Execution Host: Host1
CPU Number Ratio = 1.000000
CPU Number Ratio of RSG = {
RSG 0 = 1.000000
}
Memory Size Ratio = 0.000000
Memory Size Ratio of RSG = {
RSG 0 = 0.000000
}
Eco Status = {
Status = EXCLUDED
State Transition Time = 2017-06-20 10:49:36
Exclude Reason = HW_FAILURE
DC-OFF Times (Day) = 0
DC-OFF Times (ACCUM) = 0
}
Hardware Failure = {
Status = CPUERR
}
Execution Host: Host2
CPU Number Ratio = 1.000000
CPU Number Ratio of RSG = {
RSG 0 = 1.000000
}
Memory Size Ratio = 0.000000
Memory Size Ratio of RSG = {
RSG 0 = 0.000000
}
Eco Status = {
DC-OFF Times (Day) = 0
DC-OFF Times (ACCUM) = 0
}
Hardware Failure = {
Status = EXCLUDED
Exclude Reason = VE_DEGRADATION
VE Degradation = YES
}
```

(2)　関連説明書

NQSV 利用者の手引［JobManipulator編］4.22 実行ホストの詳細情報表示機能

2.1.4　実行ホスト数拡大

(1)　機能

１つのバッチサーバで管理する最大実行ホスト数を2048から10240に拡大しました。

(2)　関連説明書

NQSV 利用者の手引［導入編］1.2. NQSVの構成要素

NQSV 利用者の手引［操作編］第16章 制約事項

2.1.5　省電力停止回数制限の拡大

(1)　機能

1 日あたりの動的省電力運用機能によるノード停止回数を12から200に拡大しました。

(2)　関連説明書

NQSV 利用者の手引［JobManipulator編］4.16.2.（5）省電力停止回数制限

2.2　主な変更点

2.2.1　デーモン管理

(1)　機能

Linuxに適合するために、NQSVデーモンの管理がinit.dコマンドから、systemctlコマンドに変わりました。

(2)　関連説明書

NQSV 利用者の手引［管理編］1. Unit の管理

2.2.2　インストールパスの変更

(1)　機能

Linuxに適合するために、各コンポーネントのインストールパス、ディレクトリ構成を変更しました。

ex： /usr/sbin/nqsII/nqsd　→　/opt/nec/nqsv/sbin/nqs_bsvd

ex： /etc/nqsII　→　/etc/opt/nec/nqsv

ディレクトリ構成

ディレクトリ	説明
/opt/nec/nqsv/bin	コマンドのバイナリファイル置き場
/opt/nec/nqsv/sbin	管理者向けコマンド、およびデーモンのファイル置き場
/opt/nec/nqsv/sbin/systemd_prog	起動・停止用スクリプト置場
/opt/nec/nqsv/etc	コンフィグファイル置き場
/var/opt/nec/nqsv	データベース、ログファイル置き場
/opt/nec/nqsv/include	ヘッダファイル置き場
/opt/nec/nqsv/lib64	共有ライブラリ置き場
/opt/nec/nqsv/man	man データ
/usr/local/lib/systemd/system	ユニット定義ファイル置場

(2)　関連説明書

なし

2.2.3　ライセンス管理

(1)　機能

ライセンス管理方法を変更しました。

(2)　関連説明書

NQSV 利用者の手引 [導入編] 2.2 インストール

NQSV 利用者の手引 [管理編] 2.3.10. ライセンスの取得

2.2.4　スケジューリングプライオリティの設定コマンドを変更

(1)　機能

スケジューリングプライオリティの設定コマンドを変更しました。

ex： set plugin xxx　→　set priority xxx

(2)　関連説明書

NQSV 利用者の手引［JobManipulator編］3.1.3 スケジューリングプライオリティ

NQSV 利用者の手引［JobManipulator編］6. コマンドリファレンス

2.2.5　SX シリーズ（SUPER-UX）固有機能

バッチサーバ

(1)　CPU休止による省電力運用

qmgr(1M)の以下のオプションは利用できません。

- set batch_server cpu_eco_mode
- help set batch_server cpu_eco_mode

qstat(1) の以下の項目は表示されません。

- qstat -Bf
  CPU Eco Mode
- qstat -Eft
  CPU Status

(2)　マルチノード資源グループ

qmgr(1M)の以下のサブコマンドは利用できません。

- create node_group type=multinode

(3)　NQSII-BSV Agent パッケージ

(4)　SUPER-UX カーネルパラメータ

qalter(1) の以下のオプションは削除しました。

- qalter -K <parameter-name>

(5)　ジョブ実行中のジョブマイグレーション

qmgr(1M)の以下のサブコマンドは利用できません。

- set execution_queue reserve_id
- set execution_queue per_job gpid_number_limit
- set execution_queue standard per_job gpid_number_limit
- set global_queue reserve_id
- set global_queue per_job gpid_number_limit
- set global_queue standard per_job gpid_number_limit
- set execution_queue restart_option
- set global_queue restart_option
- delete execution_queue restart_option
- delete global_queue restart_option

- help set execution_queue reserve_id
- help set execution_queue per_job gpid_number_limit
- help set execution_queue standard per_job gpid_number_limit
- help set global_queue reserve_id
- help set global_queue per_job gpid_number_limit
- help set global_queue standard per_job gpid_number_limit
- help set execution_queue restart_option
- help set global_queue restart_option
- help delete execution_queue restart_option
- help delete global_queue restart_option

(6)　MPI/SX サポート

(7)　SUPER-UX 固有の実行ホスト情報表示

qstat(1)の以下の項目は表示されません。

- qstat -Ef
  Reserve ID
  RSG Resource Information
  RSG Average Information
- qstat -Sf
  RSG Number
  RSG Resource Information
  RSG Average Information

JobManipulator

(1)　マルチノードMPI/SXジョブ用HW資源制限の設定

- コンフィグファイルのマルチノードMPI/SXジョブ用HW資源の制御機能を使用に関するパラメータ（JID_CONTROL）は利用できません。

(2)　マルチノード資源グループの設定と表示コマンド

- sstat(1) のマルチノード資源グループを表示するオプション-Gは利用できません。
- smgr(1M)のマルチノード資源グループのGBC/GCR閾値を設定する下記のサブコマンドは利用できません。
  set node_group multimode_resource

(3)　smgr(1M)のマルチノードMPI/SXリクエストを集中させるためのクラスタ集中アサイン機能のためのサブコマンドは利用できません。

set queue cluster_concentration_assign

(4)　SX特有のAC電源共有ノードの優先アサインポリシー

smgr(1M)の下記のサブコマンドにac_power_shareは設定できません。

set assign_policy_priority

(5)　SX特有の拡張クラスタ向けのIXS-B列ノード優先アサインポリシー

コンフィグファイルのEXTENDED_CLUSTERパラメータは利用できません。

第3章　　NQSII R3.00 からの差分

3.1　新規機能

3.1.1　新規 MPI サポート

(1)　機能

MVAPICH2 をサポートしました。

(2)　関連説明書

NQSV 利用者の手引［管理編］10.3. MVAPICH2環境の設定

3.1.2　グループ指定実行機能のサポート

(1)　機能

リクエストを投入時に指定したグループ権限で実行することが可能です。

(2)　関連説明書

NQSV 利用者の手引［管理編］11. リクエストのグループ

3.1.3　グループ毎・ユーザ毎の制限をサポート

(1)　機能

グループ名またはユーザ名を指定した制限が設定可能です。

(2)　関連説明書

NQSV 利用者の手引［管理編］12. グループ毎・ユーザ毎の制限

3.1.4　GPU リソース管理機能サポート

qsub(1) に GPU オプションを追加

(1)　機能

投入時に GPU数を要求できるようオプションを追加しました。

(2)　関連説明書

NQSV 利用者の手引［管理編］13. VEおよびGPU対応

実装 GPU 数追随機能

(1)　機能

障害や復旧等により、実行ホストの実装GPU 数が変化した場合、最新のGPU数に基づいてスケジューリングを行います。

(2)　関連説明書

NQSV 利用の手引 [JobManipulator編] 4.10 実装CPU/GPU数追随機能

3.1.5　ソケットスケジューリング機能サポート

(1)　機能

NUMA アーキテクチャのスカラーマシン（Linux）を実行ホストとして使用する場合、ジョブに対して、最適なリソース（CPU 数、メモリ）の割り当てを行うこと（ソケットスケジューリング）ができます。 また、Linux のCPUSET 機能と連携して、リソース分割することも可能です。

(2)　関連説明書

NQSV 利用者の手引 [管理編] 19. ソケットスケジューリング

3.1.6　カスタムリソース情報サポート

(1)　機能

カスタムリソース機能とは、定義されたカスタムリソース情報に基づき、スケジューリングにて同時に使用するカスタムリソースの利用量を制御する機能です。

(2)　関連説明書

NQSV 利用者の手引 [管理編] 18. カスタムリソース機能

3.1.7　事前予約機能強化

(1)　機能

事前予約機能は、事前に指定区間のリソースを予約することで、ジョブを実行させない保守用予約区間やユーザのリクエストを確実に実行させるための予約区間を確保するための機能です。

事前予約機能に以下の機能を追加しました。

- 予約区間のアカウンティング

- 利用グループを指定した予約区間の作成
- 緊急キュー以外のキューへの予約区間の作成
- ヘルスチェックとクリーンアップ
- テンプレート指定の予約

(2)　関連説明書

NQSV 利用者の手引［JobManipulator編］4.7 事前予約機能

3.1.8　ランリミット設定

(1)　機能

ランリミットは、同時実行可能なリクエスト数の制限値です。ランリミット設定に以下のオプションを追加しました。

- ユーザ個別の同時実行リクエスト数制限
- グループ単位の同時実行リクエスト数制限
- 同時実行CPU台数制限

(2)　関連説明書

NQSV 利用者の手引［JobManipulator編］2.7.1 ランリミット設定

3.1.9　フックスクリプト機能

(1)　機能

フックスクリプト機能とは、リクエストが特定の状態に遷移した際に、バッチサーバホスト上で管理者が定義した任意のスクリプト（フックスクリプトと呼ぶ）を実行する機能です。

(2)　関連説明書

NQSV 利用者の手引［管理編］14. フックスクリプト機能

3.1.10 ユーザプリ・ポストスクリプト機能

(1)　機能

ユーザプリ・ポストスクリプト機能とは、ジョブの実行前（PRE-RUNNING）または実行後（POST-RUNNING）に、リクエスト投入時に指定した任意のスクリプト（ユーザPP スクリプトと呼ぶ）を、実行する機能です。

(2)　関連説明書

NQSV 利用者の手引［管理編］15. ユーザプリ・ポストスクリプト機能

3.1.11 初回ステージイン時間の設定機能

(1) 機能

ファイルステージングを行うリクエストが、スケジューラマップの先頭付近にアサインされた場合、ステージングが間に合わず開始予定時刻が取り消される可能性があります。それを回避するため、初回ステージイン時間をスケジューラ単位で設定可能としました。

(2) 関連説明書

NQSV 利用者の手引［JobManipulator編］4.20 初回ステージイン時間の設定機能

3.1.12 プレステージング機能

(1) 機能

ステージングなしにリクエストのアサインができる機能をサポートしました。リクエストのアサインやエスカレーション時に多数のリクエストが同時にステージングすることによるファイルシステムへの負荷を低減することが可能です。また、リクエストがアサインされてから実行開始までのステージング回数を減らすことが可能となります。

(2) 関連説明書

NQSV 利用者の手引［JobManipulator編］4.21 プレステージング機能

3.1.13 障害検知・電源制御

(1) 機能

NQSV では、実行ホスト外から実行ホストの障害を検知する機能と、実行ホストの電源制御による省電力機能をサポートしました。

(2) 関連説明書

NQSV 利用者の手引［管理編］20. 障害検知・電源制御

3.1.14 冗長化機能

(1) 機能

バッチサーバ、アカウンティングサーバ、JobManipulator の各コンポーネントを二重化（冗

長化）し、NQSV システムをダウンさせることなく、継続稼働させることができます。

(2)　関連説明書

NQSV 利用者の手引き［管理編］21. 冗長化機能

3.1.15 OpenStack と連携したプロビジョニング環境

(1)　機能

OpenStackと連携し、実行ホスト内のジョブ実行環境を動的に構成することが可能です。

(2)　関連説明書

NQSV 利用者の手引［管理編］16. OpenStack と連携したプロビジョニング環境

3.1.16 Docker と連携したプロビジョニング環境

(1)　機能

コンテナ型仮想化が可能なソフトウェアDocker と連携し、ジョブを実行ホスト内の隔離したシステム（コンテナ）上で実行できます。

(2)　関連説明書

NQSV 利用者の手引［管理編］17. Docker と連携したプロビジョニング環境

3.1.17 SCACCT 機能を NQSV に統合

SCACCTによるアカウンティング・予算管理機能をNQSVのアカウンティングサーバとして統合しました。SCACCTとの差分は以下の通りです。

(1)　各モジュールの対応関係

SCACCT	NQSV
トップサーバ	アカウンティングサーバ
中間サーバー	不要
エージェント	不要
モニタ	アカウンティングモニタ
CUI	AUI

(2)　管理するアカウンティング情報

	SACCT	NQSV
リクエストアカウント	Yes	Yes
ジョブアカウント	Yes	Yes
プロセスアカウント	Yes	No
予約アカウント	No	Yes

(3)　課金レートの設定単位の変更

NQSII R3.00：　ノード単位（SCACCTのエージェント単位)、キュー単位

NQSII R4.00・NQSV R1.0x：キュー単位、テンプレート単位

リクエスト投入時に指定したテンプレートに対して以下のように課金レートを設定します。

```
# subedit add -t template_name:CPU=0.1,MEM=0.234,DEC=0.5,ACT=0.5
rate data (template_name) add(or update) done
```

(4)　予算種別の優先順位

課金対象とする予算種別の優先順位（アカウントコード/ユーザ/グループ）をアカウンティングサーバの設定ファイル（NQSII R4.00の場合は/etc/nqsII/asvd.conf、NQSV R1.0xの場合は/etc/opt/nec/nqsv/asvd.conf）のSBU_ORDERパラメータで設定できるようになりました。

```
# cat /etc/opt/nec/nqsv/asvd.conf
#RECV_PORT_FOR_ACCT=6542
#ALLOW_CLIENTS=
SBU_CHECK=ON
#RECV_PORT_FOR_SBU=4595
SBU_ORDER=AGU
#LOG_FACILITY=LOG_LOCAL0
#ACCT_DIR=/var/opt/nec/nqsv/asv/master
#LOCK_DIR=/var/opt/nec/nqsv/asv/master
```

(5)　予定課金額

NQSII R4.00・NQSV R1.0Xよりリクエストおよびリソース予約区間に対して予定として課金する予定課金額を表示・変更する機能を追加しました。

下記のESTIMATEは、予定課金額になります。

```
# budgetedit
============================================================
USER                REMAIN        ESTIMATE         INITIAL
============================================================
usr1              11223.41           10.00        12245.00
usr2            1395382.88            0.00      1399445.00
usr3             126555.98            0.00       126555.98
============================================================
GROUP               REMAIN        ESTIMATE         INITIAL
============================================================
grp5                 0.00            0.00         1111.00
grp4                 0.00            0.00     19874344.00
============================================================
ACCOUNT             REMAIN        ESTIMATE         INITIAL
============================================================
acct1                0.00            0.00         1111.00
acct2                0.00            0.00     19874344.00
```

(6)　関連説明書

NQSV 利用者の手引［アカウンティング・予算管理編］

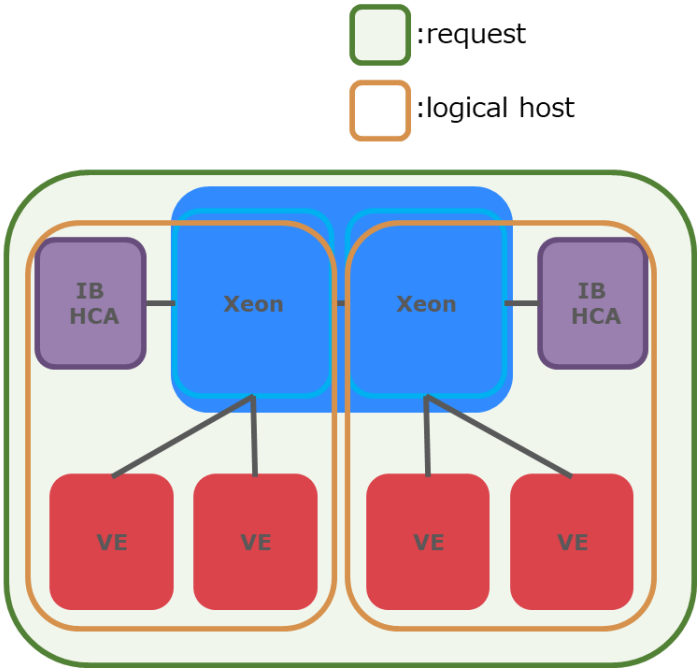
付録 A　NQSV リクエストの投入方法

A.1　VEを使用したリクエスト

VEを4台使用するリクエストのジョブスクリプトの例を以下に示します。

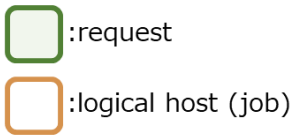
```
#PBS -b 2
#PBS -T necmpi
#PBS -l elapstim_req=300
#PBS --cpunum-lhost=2
#PBS --memsz-lhost=10gb
#PBS --use-hca=1
#PBS --venum-lhost=2

mpirun -nn 2 -np 8 ./ve.out
```



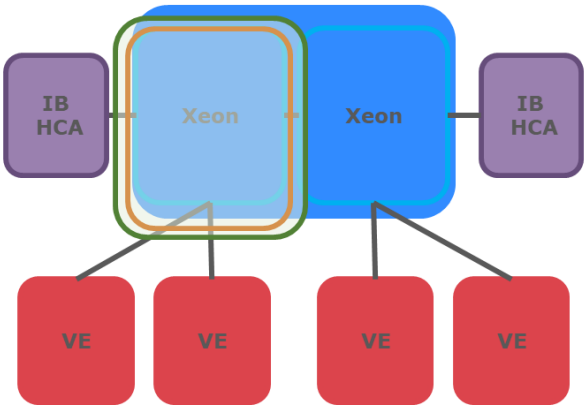
A.2　x86 を使用したリクエスト

VHのx86 CPUのリソースのみを使用するリクエストのジョブスクリプトの例を以下に示します。



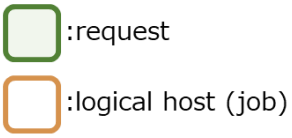

```
#PBS -b 1
#PBS -l elapstim_req=300
#PBS --cpunum-lhost=8
#PBS --memsz-lhost=10gb

export OMP_NUM_THREADS="8"
./vh.out
```



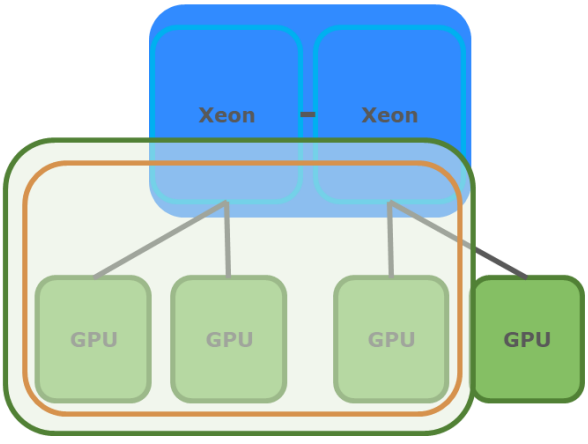
A.3　GPUを使用したリクエスト

GPGPUのリソースを使用するリクエストのジョブスクリプトの例を以下に示します。




```
#PBS -b 1
#PBS -l elapstim_req=300
#PBS --cpunum-lhost=8
#PBS --gpunum-lhost=3

./cuda.out
```



A.4　NQSVの資源制限一覧

リクエスト		
NQSV/NQSII	説明	
-l elapstim_req	経過時間制限値	
論理ホスト		
NQSV	NQSII	説明
--cpunum-lhost	-l cpunum_job	CPU 台数制限値 SX-Aurora TSUBASA の場合、ジョブで使用するVHのCPU台数です
--cputim-lhost	-l cputim_job	CPU 使用時間制限値 SX-Aurora TSUBASA の場合、ジョブで使用するVHのCPU使用時間です
--gpunum-lhost	-l gpunum_job	GPU 台数制限値
--memsz-lhost	-l memsz_job	メモリサイズ制限値 SX-Aurora TSUBASA の場合、ジョブで使用するVHのメモリサイズです
--venum-lhost	------	VE ノード数制限値
--vmemsz-lhost	-l vmemsz_job	仮想メモリサイズ制限値 SX-Aurora TSUBASA の場合、ジョブで使用するVHの仮想メモリサイズです
-l socknum_job	-l socknum_job	ソケット数制限値
プロセス		
NQSV/NQSII	説明	
-l coresz_prc	コアファイルサイズ制限値	
-l cputim_prc	CPU 使用時間制限値 SX-Aurora TSUBASA の場合、VHで実行するプロセスのCPU台数です	
-l datasz_prc	データサイズ制限値 SX-Aurora TSUBASA の場合、VHで実行するプロセスのデータサイズです	
-l filenum_prc	同時オープンファイル数制限値	
-l filesz_prc	ファイルサイズ制限値	
-l memsz_prc	メモリサイズ制限値	

	SX-Aurora TSUBASA の場合、VHで実行するプロセスのメモリサイズです
-l stacksz_prc	スタックサイズ制限値 SX-Aurora TSUBASA の場合、VHで実行するプロセスのメモリサイズです
-l vmemsz_prc	仮想メモリサイズ制限値 SX-Aurora TSUBASA の場合、VHで実行するプロセスの仮想メモリサイズです

付録 B　ACCT 全項目

NQSII R3.00 (SX-ACE, x86)、R4.00、NQSV の比較表です。

B.1　リクエストアカウント

項目名	内容	NQSII R3.00 SX	NQSII R3.00 x86	NQSII R4.00 x64	NQSV R1.0X
REQUEST-ID	リクエスト ID	○	○	○	○
REQUEST-NAME	リクエスト名	○	○	○	○
USER NAME	投入ユーザ名	○	○	○	○
GROUP NAME	グループ名	○	○	○	○
ACCOUNT CODE	アカウントコード	○	○	○	○
QUEUE NAME	投入キュー名	○	○	○	○
QUEUED TIME	投入時刻	○	○	○	○
START TIME	実行開始時刻	○	○	○	○
END TIME	実行終了時刻	○	○	○	○
CPU (SECS)	CPU 消費時間(system + user) (sec.)	○	○	○	○
REAL (SECS)	経過時間 (sec.) (*1)	○	○	○	○
REQUEST PRTY	リクエスト優先度	○	○	○	○
NICE	ナイス値	○	○	○	○
TIME SLICE	タイムスライス値	○	○	○	
REQELAPS TIME(S)	経過時間制限値 (sec.)	○	○	○	○
REQCPU TIME(S)	要求 CPU 時間 (sec.)	○	○	○	○
REQCPU NUM	要求 CPU 数	○	○	○	○
REQMEM SIZE(K)	要求メモリ量 (KB)	○	○	○	○
REQGPU NUM	要求 GPU 数			○	○
IO (BLOCKS) MFF	MFF の I/O ブロック数	○			
IO (BLOCKS) SCD	SCSI ディスクの I/O ブロック数	○			
IO (BLOCKS) SMT	SCSI テープの I/O ブロック数	○			
FLOPS	FLOPS 値	○			
CONCURRENT FLOPS	コンカレント FLOPS 値	○			
H/W CHECK	H/W 障害フラグ（16 進数）	○		○	
EXIT STAT	終了ステータス	○	○	○	○
CHARS TRANSFD	転送文字数	○			
BLOCKS R/W	I/O ブロック数	○			

KCORE MIN	延べメモリ使用量 (KB * MIN)	○	○	○	○
MEAN SIZE(K)	平均メモリ使用量 (KB)	○	○	○	○
MAXMEM SIZE(K)	最大メモリ使用量 (KB)	○	○	○	○
INSTRCT (K)	命令実行数	○			
VECTOR INST(K)	ベクトル命令実行数	○			
VECTOR ELMT(K)	ベクトル要素数	○			
VEC-EXE (SECS)	ベクトル命令実行時間［秒］	○			
MAX PROC	実行プロセス数	○			
CPU RESIDENT TM(SECS)	プロセッサレジデントタイム	○			
QUE TYPE	キューのタイプ	○	○	○	○
NUM PROCS	Number of executed processes	○			
NODE NUM	実行ホスト数			○	○
JOBS	ジョブ数	○	○	○	○
SUBREQ	サブリクエスト数 (パラメトリックリクエストのみ)	○	○	○	○
FPEC(K)	浮動小数点データ実行要素数	○			
CMCC（SEC)	オペランドキャッシュミス時間	○			
BCCC(SEC)	バンクコンフリクト時間	○			
ICMCC(SEC)	命令キャッシュミス時間	○			
MNCCC(SEC)	メモリネットワーク競合時間	○			
MT-OPEN COUNTS	MT オープン回数	○			
M/S	マルチタスクフラグ	○			
RERUN COUNT	RERUN 回数	○	○	○	○
PRERUN COUNT	PRERUN 回数	○	○	○	○
MAX NTASK	最大生成物理タスク数	○			
TEMPLATE NAME	テンプレート名 (*4)			○	○
crname(*2)	カスタムリソースの消費量			○	○
REQVE NUM	要求 VE ノード数 (*3)				○
RSVVE NUM	確保 VE ノード数 (*3)				○
VE CPU(S)	VE の CPU 消費時間 [SEC] (*3)				○
VE KCORE MIN(K)	VE の延べ使用メモリ [KB * MIN] (*3)				○
VE MEAN SIZE(K)	VE の平均メモリ使用量 [KB] (*3)				○
VE MAXMEM SIZE(K)	VE の最大メモリ使用量 [KB] (*3)				○

*1 REALはリクエストのRUNNING状態の時間です。
*2 crnameはカスタムリソース名になります。
*3 これらの項目は実行ホストがSX-Aurora TSUBASAシステムの場合のみ有効です。
*4 これらの項目は実行ホストがSX-Aurora TSUBASAシステムの場合は利用できません。

B.2 ジョブアカウント

項目名	内容	NQSII R3.00 SX	NQSII R3.00 x86	NQSII R4.00 x64	NQSV R1.0X
JOB ID	ジョブ ID	○	○	○	○
REQUEST-ID	リクエスト ID	○	○	○	○
REQUEST NAME	リクエスト名	○	○	○	○
USER NAME	投入ユーザ名	○	○	○	○
GROUP NAME	グループ名	○	○	○	○
ACCOUNT CODE	アカウントコード	○	○	○	○
HOST-NAME	実行ホスト名	○	○	○	○
QUEUE NAME	投入キュー名	○	○	○	○
QUEUED TIME	投入時刻	○	○	○	○
START TIME	実行開始時刻	○	○	○	○
END TIME	実行終了時刻	○	○	○	○
CPU (SECS)	ＣＰＵ消費時間 （system + user) (sec.)	○	○	○	○
REAL (SECS)	経過時間（sec.)	○	○	○	○
REQUEST PRTY	リクエスト優先度	○	○	○	○
NICE	ナイス値	○	○	○	○
TIME SLICE	タイムスライス	○	○	○	
REQELAPS TIME(S)	経過時間制限値（sec.)	○	○	○	○
REQCPU TIME(S)	要求ＣＰＵ時間（sec.)	○	○	○	○
REQCPU NUM	要求ＣＰＵ数	○	○	○	○
REQMEM SIZE(K)	要求メモリ量（KB)	○	○	○	○
IO (BLOCKS) MFF	MFF の I/O ブロック数	○			
IO (BLOCKS) SCD	SCSI ディスクの I/O ブロック数	○			
IO (BLOCKS) SMT	SCSI テープの I/O ブロック数	○			
FLOPS	FLOPS 値	○			
CONCURRENT FLOPS	コンカレント FLOPS 値	○			
H/W CHECK	H/W 障害フラグ(16 進数)	○		○	
EXIT STAT	終了ステータス	○	○	○	○
CHARS TRANSFD	転送文字数	○			
BLOCKS R/W	I/O ブロック数	○			
KCORE MIN	延べメモリ使用量（KB * MIN)	○	○	○	○
MEAN SIZE(K)	平均メモリ使用量（KB)	○	○	○	○

MAXMEM SIZE(K)	最大メモリ使用量 (KB)	○	○	○	○
INSTRCT (K)	命令実行数	○			
VECTOR INST(K)	ベクトル命令実行数	○			
VECTOR ELMT(K)	ベクトル要素数	○			
VEC-EXE (SECS)	ベクトル命令実行時間［秒］	○			
MAX PROC	ジョブ内で同時に存在した最大プロセス数	○			
CPU RESIDENT TM(SECS)	プロセッサレジデントタイム	○			
QUE TYPE	キューのタイプ	○	○	○	○
WAIT TIME(SEC)	待ち時間（実行開始予定時刻と実行開始時刻の差）	○	○	○	○
NUM PROCS	実行プロセス数	○			
FPEC(K)	浮動小数点データ実行要素数	○			
CMCC	オペランドキャッシュミス時間	○			
BCCC(SEC)	バンクコンフリクト時間	○			
ICMCC(SEC)	命令キャッシュミス時間	○			
MNCCC(SEC)	メモリネットワーク競合時間	○			
MT-OPEN COUNTS	MT オープン回数	○			
M/S	マルチタスクフラグ	○			
MAX NTASK	最大生成物理タスク数	○			
REQVE NUM	ジョブの割当 VE ノード数 (*1)				○
RSVVE NUM	ジョブの確保 VE ノード数 (*1)				○
VE CPU(S)	VE の CPU 消費時間の合計 [SEC] (*1)				○
VE KCORE MIN(K)	VE の延べ使用メモリ [KB * MIN] (*1)				○
VE MEAN SIZE(K)	VE の平均メモリ使用量 [KB] (*1)				○
VE MAXMEM SIZE(K)	VE の最大メモリ使用量 [KB] (*1)				○
VE REQ NODELIST	ジョブの割当 VE ノード番号のリスト (*1)				○
VE USE NODELIST	ジョブの使用 VE ノード番号のリスト (*1)				○
VE RSV NODELIST	ジョブの確保 VE ノード番号のリスト (*1)				○
*1 これらの項目は実行ホストがSX-Aurora TSUBASAシステムの場合のみ有効です。					

B.3　課金管理

課金機能の有効・無効の設定

	SCACCT／アカウンティングサーバの設定	バッチサーバの設定
NQSII R3.00 SX/x86	SCACCT の各コンポーネントの設定ファイルに SBU_CHECK パラメータにて設定	・SCACCT サーバーの設定 set batch_server **scacct_server** ・予算超過チェックの設定 set batch_server **budget_check**
NQSII R4.00 x64	アカウンティングサーバの設定ファイルの 1 か所だけで設定	・予算管理サーバーについて SCACCT に加えて NQSII 独自のアカウンティングサーバを追加しました。どちらを使うかを下記のコマンドで選択できるようにしました。 set batch_server **acct_func = { scacct | nqs_acct}** ・NQSII 独自のアカウンティングサーバについての設定を新設しました。 サーバーの設定：set batch_server **acct_server** 予算超過チェックの設定： set batch_server **nqs_budget_chk**
NQSV R1.0X	NQSII R4.00 と同じ	・予算管理のサーバーの選択機能を廃止し、NQSV アカウンティングサーバのみとしました。

課金レート

(1)　課金項目の変更点は下表のとおりです。

項目名	内容	NQSII R3.00 SX	NQSII R3.00 x86	NQSII R4.00 x64	NQSV R1.0X
CPU	CPU 消費時間一秒当たりの金額	○	○	○	○
MEM	単位メモリ使用量(1KB × 分)当たりの金額	○	○	○	○
TRNSFR	転送文字数 1 キロバイト当たりの金額	○			
IO	I/O ブロック数(1 ブロック =4096 バイト)当たりの金額	○			
INSTRUNCTION	命令実行数 1000 命令当たりの金額	○			
VECTOR	ベクトル命令実行数 1000 命	○			

項目名	内容	NQSII R3.00 SX	NQSII R3.00 x86	NQSII R4.00 x64	NQSV R1.0X
	令当たりの金額				
VELEMENT	ベクトル要素数 1000 命令当たりの金額	○			
PROCESS	1 プロセス数当たりの金額	○			
JOB	1 ジョブ当たりの金額	○	○	○	○
MTOCNT	MT オープン回数 1 回当たりの金額	○			
VECCPU	ベクトル命令実行時間 1 秒当たりの金額	○			
FLOPEC	浮動小数点データ実行要素数 1 要素数当たりの金額	○			
DKIOBLK	通常ディスク I/O ブロック数 1 ブロック当たりの金額	○			
ADKIOBLK	アレイディスク I/O ブロック数 1 ブロック当たりの金額	○			
MFFIOBLK	MFF ディスク I/O ブロック数 1 ブロック当たりの金額	○			
MASSDPSIOBLK	マスターデータプロセッシングシステムの I/O ブロック数 1 ブロック当たりの金額	○			
QTIOBLK	1/4" CGMT の I/O ブロック数 1 ブロック当たりの金額	○			
HCTIOBLK	1/2" CGMT の I/O ブロック数 1 ブロック当たりの金額	○			
DTIOBLK	DAT の I/O ブロック数 1 ブロック当たりの金額	○			
ETIOBLK	8mm CGMT の I/O ブロック数 1 ブロック当たりの金額	○			
HTIOBLK	1/2" MT の I/O ブロック数 1 ブロック当たりの金額	○			
SCSIDKIOBLK	SCSI ディスク I/O ブロック数 1 ブロック当たりの金額	○			
SCSIMTIOBLK	SCSI MT の I/O ブロック数 1 ブロック当たりの金額	○			
IMTIOBLK	IMT の I/O ブロック数 1 ブロック当たりの金額	○			
HMTIOBLK	HMT の I/O ブロック数 1	○			

項目名	内容	NQSII R3.00 SX	NQSII R3.00 x86	NQSII R4.00 x64	NQSV R1.0X
	ブロック当たりの金額				
GPUNUM	経過時間 1GPU1 秒あたりの金額			○	○
ELAPSE	経過時間 1 ジョブ 1 秒あたりの金額			○	○
RESERVE	リソース予約区間 1 ノード 1 秒あたりの金額			○	○
DEC	宣言量に対する重みづけを指定します。			○	○
ACT	実績量に対する重みづけを指定します。			○	○
PRI_MAX	優先度が最大の場合の重みづけを指定します。			○	○
PRI_MIN	優先度が最小の場合の重みづけを指定します。			○	○
crname (*1)	カスタムリソースの消費量 1 あたりの金額			○	○
REQVE	経過時間 1VE 1 秒当たりの金額 課金額の計算において要求 VE ノード数を使用します。(*2)				○
RSVVE	経過時間 1VE 1 秒当たりの金額 課金額の計算において確保 VE ノード数を使用します。(*2)				○
*1 crname はカスタムリソース名になります。 *2 これらの項目は実行ホストが SX-Aurora TSUBASA システムの場合のみ有効です。					

付録 C　改版履歴

C.1　発行履歴一覧表

2019 年　10 月　初版

C.2　追加・変更点詳細

SX-Aurora TSUBASA システムソフトウェア

NEC Network Queuing System (NQSV)
移行ガイド

2019年10月 初版

日本電気株式会社

東京都港区芝五丁目７番１号

TEL(03)3454-1111（大代表）