SEVENTH FRAMEWORK PROGRAMME

FP7-ICT-2013-10

DEEP-ER

DEEP Extended Reach

Grant Agreement Number: 610476

D8.4

Aurora Blade DEEP-ER Booster Prototype Operations Manual

Approved

Version: 2.0

Author(s): I. Zacharov (Eurotech)

Contributor(s): M. Rossi (Eurotech)

Date: 04.05.2017

Project and Deliverable Information Sheet

<table>
<tr><td rowspan="7">DEEP-ER Project</td><td colspan="2">Project Ref. №: 610476</td></tr>
<tr><td colspan="2">Project Title: DEEP Extended Reach</td></tr>
<tr><td colspan="2">Project Web Site: http://www.deep-er.eu</td></tr>
<tr><td colspan="2">Deliverable ID: D8.4</td></tr>
<tr><td colspan="2">Deliverable Nature: Report</td></tr>
<tr><td>Deliverable Level:
PU *</td><td>Contractual Date of Delivery:
31 / March / 2017</td></tr>
<tr><td></td><td>Actual Date of Delivery:
31 / March / 2017</td></tr>
<tr><td colspan="2">EC Project Officer: Juan Pelegrín</td></tr>
</table>

* - The dissemination level are indicated as follows: **PU** – Public, **PP** – Restricted to other participants (including the Commission Services), **RE** – Restricted to a group specified by the consortium (including the Commission Services). **CO** – Confidential, only for members of the consortium (including the Commission Services).

Document Control Sheet

<table>
<tr><td rowspan="6">Document</td><td colspan="2">Title: Aurora Blade DEEP-ER Booster Prototype Operations Manual</td></tr>
<tr><td colspan="2">ID: D8.44</td></tr>
<tr><td>Version: 2.0</td><td>Status: Approved</td></tr>
<tr><td colspan="2">Available at: http://www.deep-er.eu</td></tr>
<tr><td colspan="2">Software Tool: Microsoft Word</td></tr>
<tr><td colspan="2">File(s):DEEP-ER_D8.4_Manual_Aurora_Blade_Prototype_V2.0-Ecapproved</td></tr>
<tr><td rowspan="4">Authorship</td><td>Written by:</td><td>I. Zacharov (Eurotech)</td></tr>
<tr><td>Contributors:</td><td>M. Rossi (Eurotech)</td></tr>
<tr><td>Reviewed by:</td><td>H.Ch.Hoppe (JUELICH), N. Eicker (JUELICH)</td></tr>
<tr><td>Approved by:</td><td>BoP/PMT</td></tr>
</table>

Document Status Sheet

Version	Date	Status	Comments
1.0	31/March/2017	Final	EC submission
2.0	04/May/2017	Approved	EC approved

Document Keywords

Keywords:	DEEP-ER, HPC, Exascale, system architecture, component design, Aurora Blade, Aurora cluster, DEEP-ER Prototype, integration

Contents

List of Figures

List of Tables

Executive summary

This deliverable is a repackaged Operations Manual [3] for the Aurora DEEP-ER Booster Prototype. We have adopted the DEEP-ER project reporting style and format and this text can be fully used when operating the machine at JSC.

This manual contains practical instructions on how to stop and start the service, as well as maintenance instructions for the environment and safety operations.

Introduction

This document presents the operations manual for the Aurora DEEP-ER Booster Prototype. This Booster forms part of the DEEP-ER System implementing the “Cluster-Booster Concept” explored in the DEEP-ER project. The Cluster part is implemented as a regular commercial-off-the-shelf (COTS) installation. For the Booster part Eurotech submitted the final design specifications in the “feasibility study” as Deliverable D8.1 of the project, followed by the components description in Deliverable D8.2. The details of the final implementation and the measured parameters of the machine are presented in D8.3 with an update provided after the final submission. This deliverable D8.4 completes the documentation with the operational instructions to run the DEEP-ER Booster prototype.

As detailed in the update to D8.3 the DEEP-ER Booster prototype installed in Juelich does not fulfill one of the requirements set up from the beginning. Namely, the installed machine runs the peripherals with the PCIe generation 2 signaling speed (up to 5 Gbit/s per lane) instead of the envisaged generation 3 speed (8 Gbit/s per lane).

The problems became apparent at the validation stage of the bring-up of the machine. A very concerted effort of analyzing the problems was undertaken. Physical instrumentation pointed to the problem area around the backplane connector. The computer modeling revealed frequency poles in the connector insertion-loss function due to remaining connector stubs. The problem could be remedied with PCB manufacturing removing the stubs. However, in the time allotted to the DEEP-ER project this remedy could not be executed for the DEEP-ER prototype. Therefore, the prototype is currently installed with PCIe generation 2 signaling.

This document follows closely on the Operations Manual supplied by Eurotech for the pilot installation of the 1st chassis at JSC [3]. However, we have omitted sections that would duplicate the description given in the other documents (e.g. D8.3) and have referred those where needed.

In summary, this manual covers the manipulation of all the user manageable resources of the DEEP-ER chassis, such as the nodes, the BMC of the nodes, BMC of the root card and the mini-computer on the root card (called the Q7). Proper procedures for startup and shutdown are described, bringing the machine to user accessible state, ready to run programs. The emphasis is on safe operation of the machine.

During the pilot installation we discovered a case of pitting-corrosion which produced a leak from several nodes (cold-plates). The investigation with corresponding coolant (water) analysis resulted in operational rules to prevent the corrosion to occur in the future. These operational rules are also part of this manual.

1 Regular Power up and Power down Procedures

The Aurora chassis has to be connected to an external 48V DC power supply. The total power dissipation is up to 8 kW through the two connectors of type Anderson SB120 as shown in Figure 1:

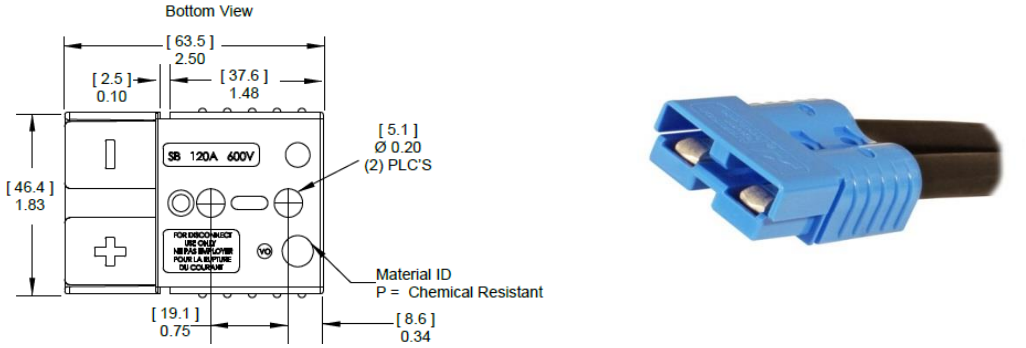


Figure 1: Power connections and Anderson connectors

The 48V DC may be supplied by the GE CP2725AC54TEZ universal (110V/220V) rectifier with the optional external controller. The external controller can be used to control the set on/off remotely. There are 4 chassis in the DEEP-ER Booster prototype installed at Juelich, and the 8 connectors are plugged into the receptacles from the existing AC/DC rectifiers already available at the installation site [4].

1.1 The First Time Power Up Procedure

To power up the system, follow the steps outlined below in the exact sequence:

1. Switch on the cooling circuit and make sure the water is flowing.
2. Verify that the Web Relay and the AC-DC are connected.
3. Verify that the Web Relay is still switched off.

   To do so, log into the deepm system as "root" user and verify that the following command:

   webrelay <webrelay address> status

   returns “off”.

4. Connect the Anderson connectors to the power supply. This way, the AC-DC is connected to the chassis.
5. Switch the Web Relay on with the command:

   webrelay <webrelay address> on

6. To switch the 1st chassis on, press the power switch on the left side of the Root card front panel (marked in red in Figure 2).

After one minute from the power on, the BMC of the root card will be booted. The BMC will request an IP address for the management port on the right side of the root card front panel (marked in yellow in Figure 2). After the boot, the controller will start to power on the devices in the chassis as follows:

- All the BMCs of the nodes will be powered on sequentially.
- All the re-timers of the root card will be powered on.
- The Q7 system on the root card is booted to user level.
- The Ethernet switch of the root card is started and configured automatically.
- All the node BMCs and the Q7 OS will send a DHCP request through the backplane interface to the external DHCP server. The DHCP server should be found from the network connected to the switch (see section 1.5.1).

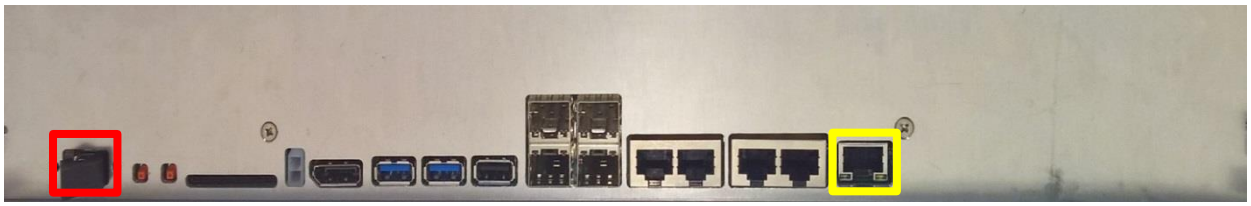


Figure 2: Root card front panel: power switch (red mark) and the management port (yellow mark).

The automatic power up procedure is now finished. The result is – all auxiliary functions (BMC, etc.) are fully functional, the KNL nodes are in standby (ready to be booted). Refer to section 1.3.2.1 for the further step on bringing up the nodes and entering the user-level service.

In general, we expect that the machine will stay connected to the mains and water supply, even when switched off. Therefore, the subsequent power up procedure will omit some of the steps specified above.

1.2 The Power Down Procedure

To power down the chassis, please follow the steps indicated below exactly in the same sequence:

- Power off the nodes via IPMI or from the OS (shutdown or poweroff commands to the OS).
- Power off the hotswap via ngpm/deeperPmTool (the tool is described in section 1.3.3).
- Power off the Q7 via software or via ngpm/deeperPmTool.
- [Optional] Switch off the Root card using the switch on the front panel.
- Power off the Web Relay.

The power off procedure is now finished.

We expect that in general the mains and the water will stay connected and does not need to be removed.

1.3 Controlling the Power State of the Nodes

To start the nodes a manual operation is required. There are several ways to control the power state of the nodes as described in the following sections.

1.3.1 IPMI

The preferred way to control the power state of the node is via IPMI.It is possible to drive the power state of the node using the following command:

```
ipmitool -H <node BMC> -U admin -P admin chassis power <status, on, off, cycle, reset, diag, soft>
```

1.3.2 The ngpm command

The ngpm is a menu-driven graphical tool (See Figure 3) which allows remote control of the power state of the following resources:

- The nodes
- The hotswap controller for each endpoint slot
- The root card OS (Q7)

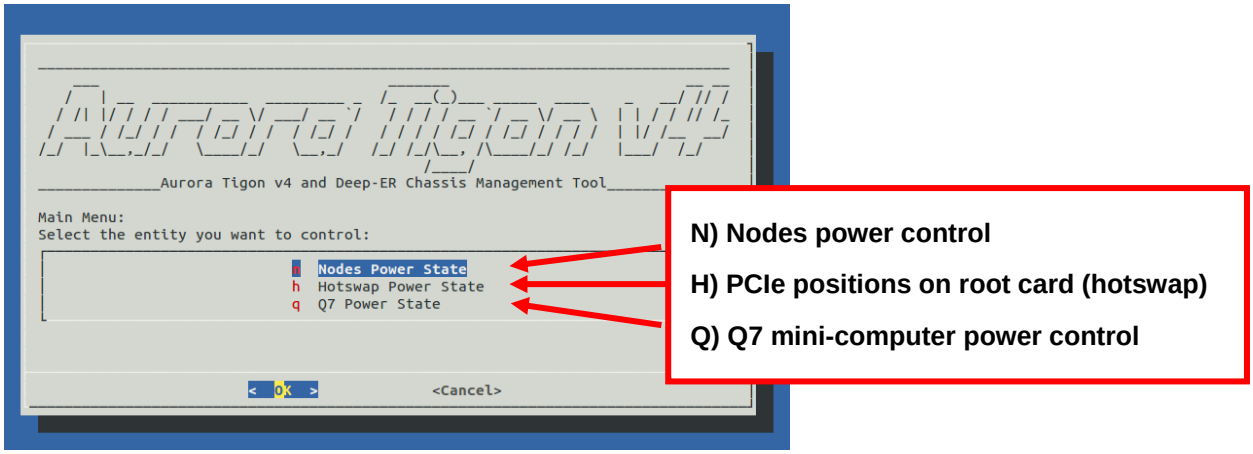


Figure 3: The ngpm tool, top level menu for action selection

The tool is provided in a .deb package and can be installed and used on every host directly connected to the BMC of the root card. The implementation could be hardware dependent and therefore encodes the hardware version in the distribution name. To install the tool use the command (for revision A root card):

dpkg -i LYNXvA_SW-<build date>-379-A.ngpm.<version number>.deb

The installation from the .deb package is primarily intended for the Q7 mini-computer on the root card. Other format distributions (for RedHat or SUSE) may be made available upon request.

The installation of this package will create the following files:

/usr/local/bin/ngpm	Graphical tool
/usr/local/bin/deeperPmTool	Command line tool
/opt/eurotech/ngpm.ini	Configuration file

The configuration file specifies the IP address/hostname (of the root card BMC) that will be used as default. The tool is preinstalled in the root card but the configuration file needs to be edited, since this value depends on the infrastructure.

To start the NGPM tool, use the command:

> ngpm

Without an explicit parameter, ngpm will use the settings specified in the configuration file.

Optionally, it is possible to specify to which root card the command should be sent changing the command line parameter.

> ngpm <BMC hostname / IP address>

1.3.2.1 Controlling the Node Power State

The second level menu shows the selection of the destination node (see **Figure 4**):

- To select the node to which the command must be sent
- To send the command to all the nodes

A third level menu will then appear, asking for the action:

- Status
- Power on
- Power off
- Reset

WARNING! When you power off a node, also the BMC of the node will be powered off. In this situation it will be no longer possible to control the power state of the node via IPMI, unless the node is started again using ngpm or deeperPmTool.

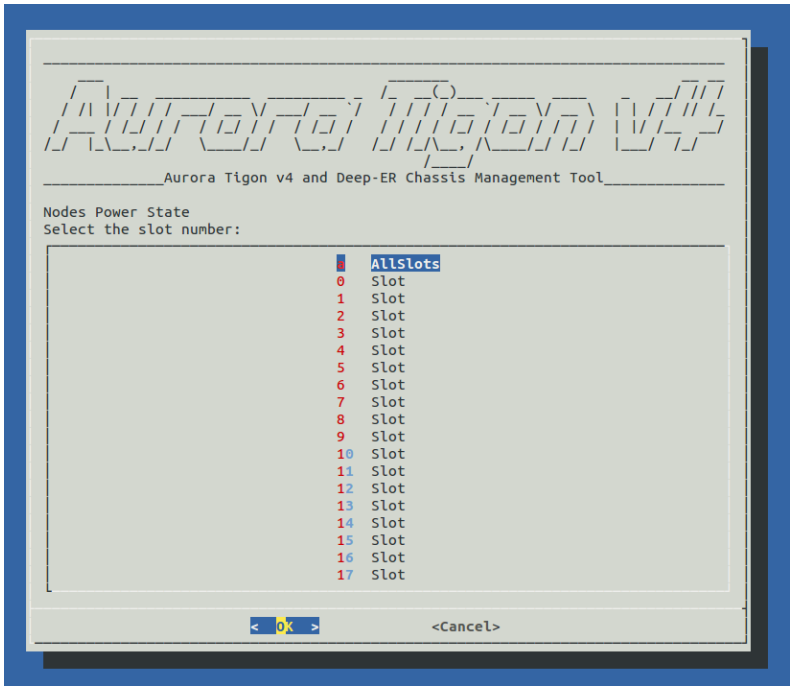


Figure 4: Slot number selection menu

Upon the power-on the nodes will automatically boot to user-level state. The OS will be loaded from the local drive or from the network resource based on the parameter in the BIOS setting of the node.

1.3.2.2 Hotswap Power State

The hotswap controller controls the power state of the PCIe slots on the root card, including the re-timer.

The upper level menu selects the destination slot for the command:

- Select the individual slot
- Select all the slots of the root card

After the selection a second level menu will allow to select the action:

- Status
- Power on
- Power off
- Reset

The PCIe peripheral slot numbering follows the node enumeration. Each node *slot* (starting from 0) is connected to *slot*2* and *(slot*2)+1* of the PCIe hotswap.

For example:

- Slot 0 → hotswap 0 and 1
- Slot 5 → hotswap 10 and 11
- Slot 17 → hotswap 34 and 35

1.3.2.3 Q7 Power State

The Q7 power state controls the OS of the root card. The upper level menu is shown in Figure 5. The following actions are possible:

Table 1. The Q7 power states

hard on	gives power to the module
hard off	removes power from the module
soft on	switches on the module (the power is already on)
soft off	switches off the module (the power is still on!)
reset	resets the module
status	gets the HARD status of the Q7 (If the host is powered OFF, the status will still be ON)

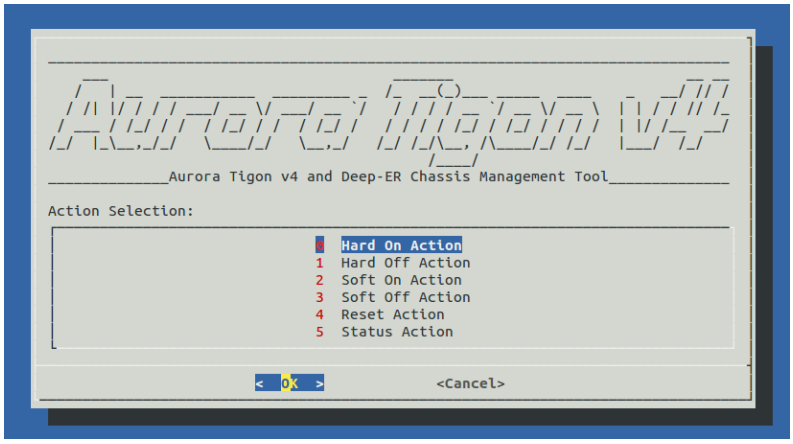


Figure 5: Q7 power state control. Selection of the action.

1.3.3 deeperPmTool

The deeperPmTool is the command line version of the ngpm tool, which allows automatizing the power management of the entities in the DEEP-ER chassis. The deeperPmTool comes with the ngpm installation and is located in /usr/local/bin

There are two ways to use the deeperPmTool:

- Interactive input
- Command line parameters

If the command is executed without parameters, the interactive version (i.e. the ngpm) is started.

The command line parameters are the following:

deeperPmTool <hostname> <entity> <action> [<slot>]

Table 2. Parameters of the deeperPmTool

<hostname>	Hostname or IP address of the root card's BMC	
<entity>	**n**	Node
	q	Q7
	h	Hotswap
<action>	**0**	Hard On
	1	Hard Off
	2	Soft On
	3	Soft Off
	4	Reset
	5	Status
<slot>	Slot number from 0 to 17(/35 for hotswap) or '**a**' for all, only in case of node action	

While for the ngpm tool the timings for the power on and power off are tuned and controlled, that is not the case for the deeperPmTool. If the power on is given to all nodes at the same time, the nodes will not boot reliably and some nodes will need to be powered up again.

WARNING! We suggest to use the "a" option for the nodes and the hotswap.
If this is not possible, or the user wants to automate the control of single nodes, we suggest delaying the sequential "hard on" command for at least 1 second. Do not send the hard on command for all the nodes in the chassis at the same time.

While the reset and the status actions are self-explanatory, he following commands of the deeperPmTool need some clarification:

hard on	power on the node/hotswap/Q7. For nodes, this command will start the BMC of the node
hard off	power off the node/hotswap/Q7. For nodes, this command will shutdown the BMC of the node.
soft on	switches on the node (power is already on)
soft off	switches off the node (power is still on!)

The "soft" action has the same effect as pressing the power button on the front panel.

The "hard" action removes power to the Q7.

If you need to power off the Q7 OS, then it is sufficient to send a "hard" action, otherwise you need to send the hard action which puts it in a standby state, following by the soft action to reboot the Q7.

The status action will return the value corresponding to the requested status:

- **0** Hard Off
- **1** Hard On – Soft off
- **2** On
- **3** Node not installed

1.3.4 Powering Nodes from the Front Panel

Verify that the external power is applied and the water is flowing – then follow up on all the steps specified for the power-up of the chassis (see Section 1.1). At this point the root card infrastructure (the BMC, Q7 and the Ethernet switch) is up and all BMCs on the nodes are up.

To power the nodes from the front panel:

- Push the power On/Off switch (the yellow light) with a suitable tool (eg. tip of a pen, the button is an insulated plastic) on the node front panel. The light will change from yellow to green and the node will start the boot sequence.

1.3.5 Powering Nodes from the Node BMC

Verify that the external power is applied and the water is flowing – then follow up on all the steps specified for the power-up of the chassis (see Section 1.1). At this point the root card infrastructure (the BMC, Q7 and the Ethernet switch) is up and all BMCs on the nodes are up.

To power the nodes from the BMC:

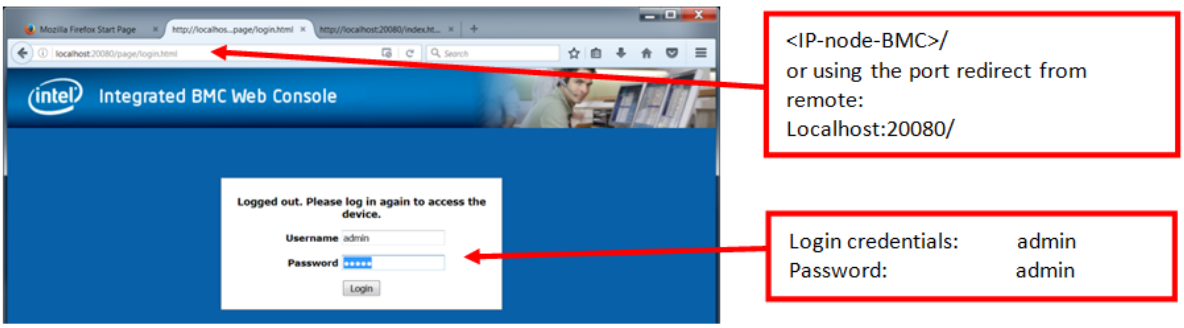
- The BMC on an individual node can be accessed from a Web browser (e.g. Mozilla Firefox)[1] by typing on the address line: <ip-of-node-bmc>/ or <hostname>s/

Note that per convention the BMC name is the node name <hostname> with extension **s**. The browser window will display the Web front page for the node.

Note: for access through the firewall port 80 must be open. Furthermore, to use the console redirect the ports 7578, 5120 and 5123 must be open for the KVM, CDROM media and Floppy/USB redirection respectively. For example, these commands

```
ssh -f -p 8000 user@gateway -L 20080:<ip-bmc-node>:80 -N
ssh -f -p 8000 user@gateway -L  7578:<ip-bmc-node>:7578 -N
ssh -f -p 8000 user@gateway -L  5120:<ip-bmc-node>:5120 -N
ssh -f -p 8000 user@gateway -L  5123:<ip-bmc-node>:5123 -N
```

will have all ports open for the remote workstation to open the Browser to the node BMC setting the address localhost:20080 and proceeding to open the console to the node on the remote workstation.



[1] Google Chrome does not work at this time due to special requirements on the Java support when bringing up the node console interface.

After the login the following screen with the system information is presented:

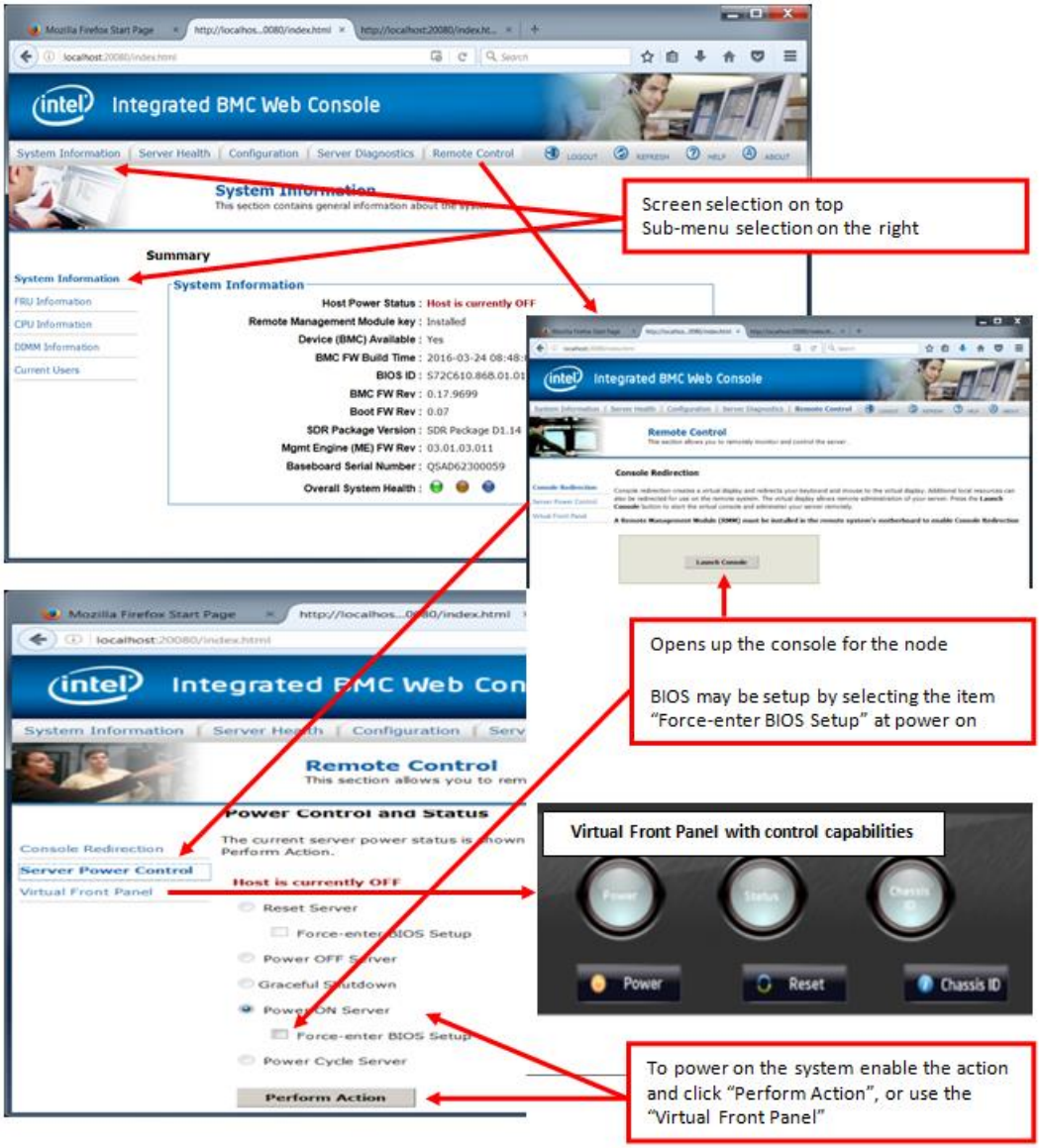


Figure 6. Control procedure using the WEB browser.

For the power, select the "Remote Control" in the screen selection menu at the top of the screen, descend to the "Server Power Control", enable the radio button "Power ON Server" and click "Perform Action".

NOTICE: The nodes can be powered through the procedure explained in this paragraph thanks to the Intel® RMM4 component. For more information, see Paragraph 1.4.

1.4 Intel® RMM4

The Intel® RMM4 is an add-on component which allows remote KVM access and control through LAN or Internet. The RMM4 is mounted on the node and allows the user to perform actions on the node itself. To learn more about the Intel® RMM4 features and its use, refer to the RMM4 user guide [5].

There is a special remark for the usage of the Serial-over-LAN (SOL) capability of the server:

WARNING! By default all nodes have the RMM4 module installed and operative. The KVM of nodes is redirected automatically through LAN. However, this disables SOL. In order to enable SOL, the user has to connect via LAN to the RMM4 and enable the SOL from the Web interface.

1.5 The Network Setup

The DEEP-ER Booster has 2 different networks: Ethernet and the EXTOLL TOURMALET interconnect (configured as a 2.5D torus).

1.5.1 Ethernet

For setting up the Ethernet network the following connections are required:

- 1 Ethernet cable for the switch of the root card (Q7, OS of the nodes and the BMC of the nodes) – to be connected to one of the connectors marked in red in Figure 7.
- 1 Ethernet cable for the BMC of the root card – to be connected to the connector circled in yellow in Figure 7.

This is valid for each chassis, therefore total of 8 Ethernet cables are needed for the DEEP-ER Booster to connect to the Juelich data center.

The 10GigE interface is fully working with root card revB hardware and this interface can also be used with DAC (Direct Attach Copper) cables. Using multiple 1GigE cables to the router in JSC is possible to increase the bandwidth of the connection. Special setting of the root card internal Ethernet switch for port bonding may be provided upon request.

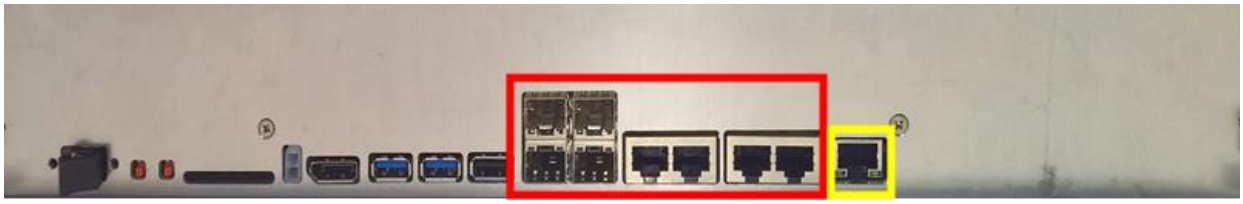


Figure 7: Root card front panel: Ethernet connections from internal switch marked red.

1.5.2 HPC network with EXTOLL TOURMALET

The settings for EXTOLL TOURMALET are provided by JUELICH and UHEI. The cabling is described in section 2.3.3 (HPC Network topology) of Deliverable D8.3 [6].

2 Liquid Cooling System Operations

To maintain the proper coolant mixture that maintains integrity of the system and prevents:

- Corrosion
- Biological contamination

The system must run the coolant mixture specified in Table 3:

Table 3. The water mixture parameters and the resulting PH and Conductivity thresholds.

Demineralized Water	**Used as base for the mixure**
Clariant Protectogen® C Aqua (anti-corrosion additive)	**min 1.5% max 2.0%** **based on final volume (v/v)** *[indirectly tested via conductivity]*
Thor ACTICIDE B40 (anti-algae/bacteria additive)	**min 0.1% max 0.2%** **based on final volume (v/v)**

The coolant has to be tested periodically to maintain the benchmark for safe operation, namely:

Parameter	**DM Water**	**Mixture**	**Comments**
pH	6 ÷ 7.5	7 ÷ 8.5	Measurement procedure: APAT-IRSA-CNR Man 29/2003 met. 2060
Conductivity mS/cm	$< 4\times10^{-3}$	4.3 ÷ 4.9	The conductivity must be measured for the required concentration to establish the reference measurement. APAT-IRSA-CNR Man 29/2003 met. 2030

The method for testing the values is specified in the comments column[2].

The coolant has to be prepared as specified above and let circulating for a few days before the first sampling must be taken to test for the chemical quality of the liquid. This first test may lead to first corrections (adding the chemicals) to attain the required parameters as specified in table above.

This first test is needed also to evaluate the effect on the coolant chemistry due to the materials involved in the cooling loop and the interaction between the coolant and the material of the infrastructure. It's recommended that Eurotech accesses the results so that, if needed, it can suggest corrective actions and confirm the list of requirements.

A new sampling and water analysis should be performed after 1 month, checking again all the specified parameters. If the results show that the coolant is compliant to the requirements and stable in time, the sampling frequency can be increased to every 3 months.

It is recommended that every three months Eurotech receives a report from the Customer regarding to the liquid analysis, in reference to values in Table 3 (pH and the conductivity) as well as for the, hardness ($CaCO_3$), metal particles, concentration of sulfides, chlorides, sulfates, etc. as much as available. The anticorrosion additive concentration and concentration of biological growth

[2] Analytical methods for water are standardized procedures. The Italian methods are similar for methods adopted in other countries of the EU. See, for example: http://www.isprambiente.gov.it/en/publications/handbooks-and-guidelines/metodi-analitici-per-le-acque-analytical-methods?set_language=en

inhibitor must also be monitored. The last parameter cannot be measured directly, but the maintenance of the pH values will facilitate the biocide activity.

To test the concentration of the anti-corrosion additive, the measurement of conductivity should be applied.

For the initial run of the DEEP-ER Booster prototype at JSC the Biocide (ACTICIDE B40) will be added after the stabilization of the initial runs, using only the corrosion inhibitor (Protectogen). As soon as possible a conductivity measurement instrument will be installed to enable continuous monitoring of the coolant quality.

3 Unforeseen Shutdowns, System Excursions and Emergencies

3.1 Critical Events

Each DEEP-ER Booster chassis is protected by the Chassis Protection System (CPS) that will automatically switch off the chassis by interrupting the main power to the chassis when critical overheating events occur. The CPS is discussed in section 4.

However, this is not sufficient to preserve the integrity of the DEEP-ER Booster rack because there are some critical events that may occur even when the chassis are off, partially off, or when electronics are already suffering and the Chassis Protection System (CPS) is not in a condition to work well and be effective. Some examples of these events are:

- Leakages or condensation phenomena that are still not impacting the behavior of electronics but anyway can become suddenly critical or slowly damage the system
- Smoke due to some burning event that is not involving the chassis itself or is occurring when the chassis are off or already in a suffering state, so that self-protection function in them cannot be reliable.

3.2 Intra-rack Sensors

To prevent such events there is a number of sensors in the rack that have to be read and managed by the infrastructure and not by the electronics in the chassis. Namely, these sensors are:

- Intra-rack ambient temperature and humidity sensors
- Intra-rack smoke sensors
- Chassis-level leakage sensor
- Intra-rack leakage sensors

3.2.1 Intra-rack Ambient Temperature and Humidity Sensors

Reading the data provided by the intra-rack temperature and humidity sensors is key to avoid condensation phenomena. Indeed, the temperature of the coolant injected in the rack has always to be higher than the dew point temperature threshold that is calculated starting from the ambient temperature and humidity (e.g. if ambient relative humidity is 60% and ambient temperature is 25°C dew point temperature is 17°C, the temperature of the coolant must thus be above 18°C to avoid intra-rack condensation).

If the coolant temperature is below the dew point the temperature of the liquid entering the rack has to be increased to avoid formation of water drops by condensation inside the chassis. For example a temporary reduction in the flow rate entering the rack can be applied, up to a complete reduction of the flow for the time needed to exit condensation regime. The interruption of the flow may be needed if the rack is not up to provide the heat to exit condensation regime.

An intra-rack ambient temperature increase must generate an alarm and the system must be shut down to prevent eventual damage until the reason for the increase is understood and eliminated. This kind of events can be related to pump failure or partial or completely clogged filters that prevent the coolant flow.

Therefore, the infrastructure should contain additional sensors:

- The flow meter sensors should be present in the pipes to measure the flow rate at the entrance to the rack, capable of raising alarms when flow rate drops below threshold.
- The differential pressure sensors over the filters or over the rack should indicate limited or blocked flow through the rack and capable raising the alarm.

In case of overheating caused by no flow or very limited flow, it is very important to avoid a sudden injection of cold coolant into the rack since this might generate thermal shocks that can damage the piping joints causing leakages inside the rack.

To manage any severe rack overheating events caused by the absence of coolant flow requires these actions:

1. Switch off the chassis
2. Remove the main power from the chassis
3. Leave the rack to cool down in a natural way

3.2.2 Intra-rack Smoke Sensors

The infrastructure management system should read the intra-rack smoke sensors and raise alarms. The alarms should lead to the immediate interruption of the main power to the chassis of the rack.

3.2.3 Chassis-level and Intra-rack Leakage Sensors

The infrastructure management system should read the leakage sensors and raise alarm leading to the shutdown of the chassis, interruption of the main power to the chassis and stop the coolant flow to the rack. Severe condensation phenomena might also lead to alarms from the leakage sensors.

3.2.4 Reading of the Sensors

We suggest to accurately prepare a specific list of actions to be undertaken in case of each critical event. Furthermore, we advise to precisely identify the thresholds and conditions that characterize each potentially dangerous event. Preferably, these actions should be performed in an automated way to be sure to have them performed in the right order and avoiding delays or mistakes caused by misinterpretation of them or human errors. This is very important to preserve the integrity of the system.

WARNING! The examples mentioned above are only some of the critical events that can occur. It is very important to read the intra-rack sensors connected to the infrastructure monitoring systems and to use the related readings together with the ones coming from the rest of the sensors present at infrastructure level in order to undertake the right actions to preserve the integrity of the system.

4 The Chassis Protection System

The chassis protection system (CPS) is a software monitor running on the Root Card BMC. The CPS is a daemon that queries temperature sensors on the nodes installed in the chassis and shuts down the full chassis if at least one node temperature exceeds the set threshold for the specified amount of time.

At the startup of the BMC the CPS reads the initialization file /conf/lynxcps.ini to configure the sensors polling frequencies and the thresholds. The configuration file is non-volatile and can be changed by system administrators to tune the settings. This file contains:

Table 4. Settings of the Chassis Protection System.

[CPS] good_event_frequency=10 warning_event_frequency=3 alert_event_frequency=1 good_ev_threshold=10 warn_ev_threshold=6 alert_ev_threshold=1	The sensor polling interval (in seconds) and thresholds (iterations) to define whether the system is in Good, Warning or Alert state. With this configuration, the algorithm changes state from Good to Warning after a minute and from Warning to Alert after three seconds if temperatures are above the warning or alert thresholds, respectively.
[sensor] name=n00t1 threshold_low_warning=5.0 threshold_high_warning=60.0 threshold_low_alert=1.0 threshold_high_alert=65.0	A sensor entry with its temperature thresholds (in Celsius degrees). The upper limits in this case are set to: Warning: 60C Alert: 65C

If at least one sensor is above the Alert threshold for 2 seconds the CPS switches the chassis off. However, if the temperature drops before the action occur the {warning, alert} state is reverted to the previous condition.

The CPS logs the warning and alert states. All messages are written to files as follows:

Alert: 10 last messages before shutdown	/conf/lynxcps.log	Available after reboot for post mortem analysis
Alert: log	/var/log/alert.log	Volatile
Warning log	/var/log/warning.log	Volatile

The changes to the /conf/lynxcps.ini are persistent across reboots. When the configuration file is changed a CPS restart is necessary to read it. This restart is triggered via the commands:

```
#/etc/init.d/lynxcps_ctl stop;  /etc/init.d/lynxcps_ctl start
```

References and Applicable Documents

[1] Technical Product Specification for Intel® Server Board S7200AP Family.
URL http://www.intelserveredge.com/assets/S7200AP_HNS7200AP_TPS_R1_0.pdf
[2] Intel® SSD DC P3700 Series Specifications.
URL http://www.intel.com/content/www/us/en/solid-state-drives/ssd-dc-p3700-spec.html
[3] DEEP-ER BOOSTER: Operation Instructions, Eurotech document 02.12.2016
[4] DEEP-ER prototype installation, D3.5
[5] Intel Remote Management Module 4, Technical Product Specification, Rev. 1.5, July 2014. Intel order number G24513-005
[6] “Aurora Blade Booster Prototype for DEEP-ER”, D8.3 and D8.3 update (31/03/2017)

List of Acronyms and Abbreviations

A

AP:	Adams Pass. Intel implementation of the KNL reference board
API:	Application Programming Interface

B

BIB	Backplane Interface Board. Electrical and signalling board to provide routing of signals between Reference board and Aurora Backplane.
BIOS:	Basic I/O system. Boot and system initialization code run before the OS starts
BLN:	Brick local network. Used to locally connect the Brick modules
BMC:	Board management controller. Used to physically monitor and manage a compute blade.
Brick:	Modular entity forming a booster node
Brick Module:	Smallest functional HW entity. Up to 6 modules are aggregated into a Brick

C

Chassis:	Mechanical entity mounted in a rack. A chassis typically aggregates multiple mechanical sub-units (here: Bricks) through a chassis level infrastructure (e.g. Backplane, power, cooling)
CPU:	Central Processing Unit

D

DAC	Direct Access Copper (DAC) – connector for 10 GigE interfaces
DEEP:	Dynamical Exascale Entry Platform
DEEP-ER:	DEEP Extended Reach: this project
DEEP-ER Global Network:	High performance network connecting Bricks, CN, NAM and other global resources to form the DEEP-ER Prototype system
DEEP-ER Prototype:	Demonstrator system for the extended DEEP Architecture
DMA:	Direct Memory Access
DFG:	Deutsche Forschungsgemeinschaft, German research organisation
DRAM:	Dynamic Random Access Memory. Typically describes any form of high capacity volatile memory attached to a CPU
DDR-4:	Interface standard to attach DRAM to a CPU
DDP	Dual Die Package

E

ECC:	Error correction code. Corrects errors in storage and transmission systems by added redundancy.
Exaflop:	1018 Floating point operations per second

Exascale:	Computer systems or Applications, which are able to run with a performance above 1018 Floating point operations per second
EXTOLL:	High speed interconnect technology for cluster computers developed by University of Heidelberg

F

FLOP:	Floating point Operation
FPGA:	Field-Programmable Gate Array, Integrated circuit to be configured by the customer or designer after manufacturing

G

H

HMC:	Hybrid Memory Cube
HMCC:	Hybrid Memory Cube Consortium
HPC:	High Performance Computing
HW:	Hardware
Hybrid Memory Cube:	Novel type of computer RAM that uses 3D packaging of multiple memory dies to increase memory capacity and number of data banks per device area. Technology is being developed by Micron Technology and backed by the Hybrid Memory Cube Consortium.
Hybrid Memory Cube Consortium:	Industry association defining HMC interfaces and facilitating HMC Integration into a wide variety of systems. Includes Samsung, Micron Technology, Open-Silicon, ARM, IBM, SK-Hynix, Altera, and Xilinx.

I

I2C:	Inter-Integrated Circuit bus. A low cost serial bus used to interconnect silicon devices. Typically used for status monitoring and configuration.
IB:	InfiniBand
Intel:	Intel Germany GmbH Feldkirchen,
I/O:	Input/Output. May describe the respective logical function of a computer system or a certain physical instantiation

J

K

KNC:	Knights Corner, Code name of a processor based on the MIC architecture. Its commercial name is Intel® Xeon Phi™.
KNL:	Knights Landing, second generation of Intel® Xeon Phi™

L

LPC:	Low Pin Count bus

M

MQTT:	Message Queue Telemetry Transport protocol
MIC:	Intel Many Integrated Core architecture

MPI:	Message Passing Interface, API specification typically used in parallel programs that allows processes to communicate with one another by sending and receiving messages
MR-IOV	Multi-root I/O virtualization. Standard to share a PCI Express endpoint between multiple hosts

N

NAND Flash memory:	Implementation of non-volatile memory used today for solid state disk.
NAM:	Network Attached Memory, nodes connected by the DEEP-ER global network to the Bricks providing shared memory buffers/caches, one of the extensions to the DEEP Architecture proposed by DEEP-ER
NIC:	Network Interface Card, Hardware component that connects a computer to a computer network
NTB:	Non-transparent bridge. A component required to connect PCI hierarchies
NVM:	Non-Volatile Memory. Used to describe a physical technology or the use of such technology in a non-block-oriented way in a computer system
NVMe:	Short form of NVM-Express
NVM-Express:	An interface standard to attach NVM to a computer system. Based on PCI Express it also standardizes high level HW interfaces like queues.

O

OpenMP:	Open Multi-Processing, Application programming interface that support multiplatform shared memory multiprocessing
OS:	Operating System

P

PA:	Physical address space. Used on hardware level to access system components.
PC:	Personal Computer
PCB:	Printed circuit board.
PCH:	Platform controller hub. Companion device to provide commodity peripherals to Intel® CPUs
PCI:	Peripheral Component Interconnect. Originally linked to a dedicated physical implementation, it now stands for a standardized way to attach and manage peripherals in computer systems
PCIe:	Short form of PCI Express
PCI Express:	Peripheral Component Interconnect Express started as an option for a physical layer of PCI using high-performance serial communication. It is today's standard interface for communication with add-on cards and on-board devices, and makes inroads into coupling of host systems. PCI Express has taken over specifications

	of higher layers from the PCI baseline specification.
PCISIG:	PCI special interest group. Industry association responsible for the development of the PCI/PCI Express standards
PCM:	Phase change memory. A technology candidate for future non-volatile memories
PFlop/s:	Petaflop, 1015 Floating point operations per second
PLX:	Provider of PCI Express system components

Q

QPACE:	Specialised supercomputer for QCD Parallel Computing on CELL processors

R

Rack:	Compartment to mechanically assemble multiple chassis to form the final computer
RAM	Random-Access Memory
RDMA:	Remote Direct Memory Access
RDIMM	Registered Dual In line Memory Module

S

SDP	Single Die Package
SKU	Shelf Keep Unit – a single logical storage item, that may consist of multiple parts
SOL	Serial Over Lan
SM-Bus:	Single-ended simple two-wire bus derived from I2C for the purpose of lightweight communication often used management of computer system components.
SSD:	Solid State Disk
SW:	Software

T

TSV	Thru Silicon Via
TLP:	Transaction layer packet. Basic packet structure to transport transactions across a PCI Express infrastructure.

U

U	Linear unit of measure, 1U = 1.75” = 44.45 mm

V

VF:	Virtual function. A functional element of a PCI endpoint
VLP RDIMM	Very Low Profile RDIMM

W

WAN:	Wide Area Network
WP:	Work Package

Operations Manual

X

x86: Family of instruction set architectures based on the Intel 8086 CPU

Z

ZITI Heidelberg: Institut für Technische Informatik Uni Heidelberg, Germany