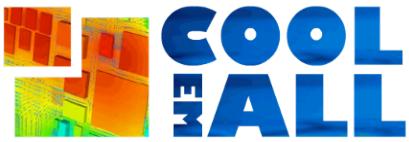

Project acronym: **CoolEmAll**

Project full title: **Platform for optimising the design and operation of modular configurable IT infrastructures and facilities with resource-efficient cooling**

D3.1 First definition of the flexible rack-level compute box with integrated cooling

Author: Micha vor dem Berge (Christmann)

Version: 1.0

Date: 27/03/2012

Deliverable Number:	D3.1
Contractual Date of Delivery:	31/03/2012
Actual Date of Delivery:	31/03/2012
Title of Deliverable:	First definition of the flexible rack-level compute box with integrated cooling
Dissemination Level:	Public
WP contributing to the Deliverable:	WP 3
Co-authors:	Micha vor dem Berge, Wolfgang Christmann (Christmann) Eugen Volk (HLRS) Assunta Napolitano (IREC), Jean-Marc Pierson, François Thiebolt, Georges Da Costa (UPS)

History			
Version	Date	Author	Comments
0.1	07/12/2011	Micha vor dem Berge (Christmann)	Initial Draft
0.2	03/02/2012	Assunta Napolitano (IREC)	First Draft for section 2.3
0.3	15/02/2012	Eugen Volk (HLRS)	First Draft for section 3
0.4	17/02/2012	Micha vor dem Berge, Wolfgang Christmann (Christmann)	First Draft for sections 1, 2.1, 2.2, 2.4, 2.5
0.5	19/02/2012	Wolfgang Christmann (Christmann)	Additions to section 2.2
0.6	20/02/2012	Assunta Napolitano (IREC)	Additions to section 2.4
0.7	24/02/2012	Assunta Napolitano (IREC), Micha vor dem Berge (Christmann)	Additions to section 2.4, Additions to section 2.5
0.8	29/02/2012	Eugen Volk (HLRS), Micha vor dem Berge (Christmann)	Additions to section 3, Additions to section 4
0.9	29/02/2012	Jean-Marc Pierson, François Thiebolt, Georges Da Costa (IRIT)	Corrections and minor additions to sections 2 and 3
0.91	12/03/2012	Assunta Napolitano, Ramon Berenguer Fornós (IREC)	Corrections and minor additions to section 2.3

0.91	19/03/2012 13/03/2012	Jean-Marc Pierson (IRIT), Ramon Berenguer Fornós (IREC)	Review of version 0.91
0.92	21/03/2012	Eugen Volk (HLRS)	Addressing reviewers' comments and update on section 3.6.5
0.93	23/03/2012	Micha vor dem Berge (Christmann)	Addressing reviewers' comments
0.94	26/03/2012	Eugen Volk (HLRS)	Addressing reviewers' comments
1.0	27/03/2012	Micha vor dem Berge (Christmann)	First version

Approval		
Date	Name	Signature
30/03/2012	Ariel Oleksiak	

Abstract

In this Deliverable we offer a first definition of a flexible rack-level compute box (ComputeBox1) with integrated cooling. This ComputeBox1 is a fully integrated solution which is designed to be built and delivered to data centres as a complete block ready to use.

In this deliverable we identify the requirements of such a ComputeBox1, such as high density, energy-efficiency, integrated cooling and integrated monitoring. These requirements lead to specific design-decisions concerning the individual parts of the ComputeBox1. For example, we decided to use the high-density RECS Cluster Computer system as the basis for the computing nodes which offers a very high density and energy-efficiency, as well as an integrated high capable monitoring- and controlling-solution. This enables us to monitor the complete Rack at a very fine granularity without a negative impacting the computing- and network-resources.

For every single key-component we analyse the state of the art from industry and actual research projects, where possible we additionally offer own solutions which go beyond. Through this sensible selected mix of state of the art, well-tested components and high potential new products or ideas we expect an innovative but stable setting.

In addition to the ComputeBox1 or rack-level compute box, this deliverable contains a definition of Data centre Efficiency Building Blocks (DEBB). A DEBB is an abstract description of a piece of hardware on different granularity levels. These granularities reach from a single node up to a complete data centre and will help users to model and simulate a virtual data centre for e.g. planning or reviewing processes.

Keywords

Rack-level compute Box, ComputeBox1, Data Centre Appliance in a Rack, Cooling, Integrated Monitoring, High Density, Energy-Efficiency, RECS, Data centre Efficiency Building Block (DEBB)

Table of Contents

List of Figures

List of Tables

1 Introduction

Most data centres are based on server racks, aligned as rows. Each rack contains a variety of servers, storage-systems and peripherical equipment. These racks are in most cases build up step by step and there is some effort to invest to integrate all components.

For building up big data centres, some vendors and data centre providers have done some efforts for new concepts of a higher integration level for the computing and storage infrastructure, mostly based on container-modules. These units are only interesting for real big data centres. For smaller companies with lower needs of IT-Infrastructure these containers are oversized.

If we want to make the IT infrastructure more resource efficient we have to look with a special focus at the smaller data centres. Google, Amazon and Facebook do big own efforts for a higher efficiency and build up complete own designed solutions. The smaller data centres are not able to proceed like this. They depend on available and affordable solutions for smaller units.

In CoolEmAll we are looking for a higher integrated rack-level compute box that should make computing as resource efficient as possible. This “ideal server rack” is defined as follows:

- High density of the integrated computing and storage capacity
- Avoiding cables as far as possible by higher integration of components and stackable components
- Integrated cooling solution that is placed very near to the hot spots of these systems
- Integrated sensoring and monitoring infrastructure that delivers exact values for heat, power consumption and other important system parameters
- Integrated controlling infrastructure that allows to adopt the system hardware in a most granular way to the needed performance

In CoolEmAll it will not be possible to build up a whole prototype, because there are some constraints of the budget. But we will define the rack level compute box in detail and build up a realistic prototype that will allow us to perform different real world scenarios in this testbed. With this real world data and additional simulations we will get conclusions for the behaviour of a fully populated compute box. Based on this we will redefine the concept of the optimal rack-level compute box.

2 Definition of the rack-level compute box

In the following sections, the rack-level compute box (ComputeBox1) will be defined in a top-down method. First, we will collect requirements to the rack-level compute box. Second, we analyse these requirements to define a first mechanical and functional draft. The draft will be described from the outer parts (physical overview, cooling) to the inner parts (components within the rack-level compute box, integrated monitoring & controlling) to give an overview of all important parts.

2.1 Requirements

In this section the requirements will be collected and evaluated. The collection of requirements is the most important step to define how the rack-level compute box should look like, what features it should have and how it can be used within the CoolEmAll project.

2.1.1 Material-efficiency

One important goal of the CoolEmAll project is the development of blueprints and tools to support building small, bespoke facilities for data centres. A bespoke facility means usually that the facility is as large as needed but as small as possible.

To be able to build the facility as small as possible it is necessary to know the requirements of the server hardware in detail. For this reason it is very helpful to have a pre-configured server environment like the here described rack-level compute box with high density. Furthermore the choice of material is of importance. For the environmental impact it is a big difference whether to choose e.g. thin or thick steal for the enclosures and also the eco-friendliness of the colours used should be considered.

2.1.2 High density

High density is a prerequisite for material-efficiency and thus is important to be mentioned as a requirement of the rack-level compute box. Having high density means running many servers on a very small area which might lead to cooling problems. This is why the servers have to be energy efficient and we need to have an integrated cooling (both are further requirements listed and explained below).

Modern Blade Centres have a density of two nodes on one rack unit (U), whereas each node usually can handle two CPUs. This means a density of CPUs/U. For the rack-level compute box we plan to have dedicated server units that can host more than 10 CPUs/U. Additionally to the server units we need high-efficient power supplies, which should have an efficiency of more than 95%

under all working conditions, and can power several server units. Integrating also storage and network infrastructure the overall density should be at least twice of today's typical Blade Centres.

2.1.3 Energy efficiency

High energy efficiency is a requirement that on the one hand was raised by the two predecessors and on the other hand is one of the main issues of this research project. High energy efficiency of servers can be reached by considering various factors. As stated in Koomey's Law by Koomey, Berard, Sanchez and Wong (2010), the energy efficiency of micro processors grows with the same rate predicted by Moore's Law, see Moore (1965). This means the energy efficiency doubles every 18 months, but there are still big differences between different server models. These differences have to be tested, analysed and only the most efficient servers will be used as the basis for the rack-level compute box.

With the previously described measures, we can provide energy-efficiency that is state of the art of today's servers. But in this project our goal is to reach a better energy-efficiency which is necessary as the overall energy consumption of Data Centres still grows significantly, the worldwide data centre electricity consumption grew up from 70.8 Billion kWh in the year 2000 to 152.2 Billion kWh in 2005 (Koomey, 2008).

Through a minimal invasive integrated monitoring infrastructure and different management approaches to be integrated in the rack-level compute box we plan to lower the energy consumption distinctly.

2.1.4 Stackable

Being stackable means two things. First, the rack-level compute box should be compatible with typical 19" server racks and thus can be integrated into existing data centres. Second, the rack-level compute box should be able to be placed multiple times in a row to not become an isolated solution.

Also the compute shelves inside the rack shall be stackable. This enables us to vary the size of the rack-level compute box and the number of integrated compute shelves to build best-suited sizes. By only integrating the really needed modules, we can support again the requirement of material efficiency and high density.

This requirement has a strong impact on the physical concept, as well as on the cooling solution.

2.1.5 (Re-)configurable

For the implementation, tests and validations of various scenarios (which will be described in CoolEmAll D6.1), we will need a highly flexible testbed system that can be configured and re-configured to meet the particular requirements. Besides our requirements to have a flexible test-bed, we will gain profit from the re-configurability in terms of energy-efficiency by trying to use always application-optimised hardware. This will be done by coupling application profiles to hardware profiles, see the section about DEBB below.

Re-configuration can happen on a physical layer, on server hardware layer and on the software layers. Therefore, we see at least requirements for the following aspects of the rack-level compute box to be (re-)configurable:

- CPU architecture (x86 32-Bit, x86 64-Bit, ARM)
- CPU speed (from low to high performance)
- Mainboards (some mainboards may support e.g. power saving modes from Linux, other don't)
- Main memory (RAM) size
- Cooling solution (e.g. side-blowing fans, rear-blowing fans, integrated cooling in the side or rear of the rack, no cooling solution)
- Sensors (at least for temperature we'll need some sensors that can be re-arranged, depending on the physical construction of the cooling)
- Operation System (In the evaluation, we will compare different types of scenarios thus we need a flexibility in booting different Operation Systems)

2.1.6 Integrated cooling

To be able to deploy a rack-level compute box into an empty room without explicit cooling, we need an integrated cooling solution in the compute box itself. An integrated cooling makes the compute box climatic neutral to the room. This is also good for rooms that are already populated with servers and the cooling solution is working to capacity.

A further advantage is that the cooling can be ideal adjusted to the density and thus the maximum electrical power that has to be cooled within a rack.

2.1.7 Integrated monitoring and controlling

The rack-level compute box should have a dedicated monitoring and controlling infrastructure which is independent from the compute network. The main advantage of such a dedicated monitoring and controlling infrastructure is to

avoid the potential overheads caused by measuring and transferring data, which would consume lots of computing capabilities. In particular in a large-scale environment this approach can play a significant role.

Therefore each compute node should be connected to an additional independent microcontroller in order to collect the measured data independent. The measurements should be gathered and published at a sufficient rate so as to ensure a proper action on the system on time. The benefits that we expect to be gained from an integrated monitoring and controlling infrastructure are numerous. We expect more knowledge about the workload on the servers, the energy consumption and the typical energy usage of specific workloads. Furthermore, we can publish this extracted information and make it available for users and for resource management systems.

If we also integrate intelligent management mechanisms and the possibility to individually control each part of the ComputeBox (such as switching off some boards, some disks, changing operation speeds of CPU and network cards, and so on), we expect high energy efficiency as already shown in the GAMES research project, see Cioara et al.

2.2 Physical Overview

In this section we provide an overview on server solutions used to build up energy efficient data centre, describing corresponding dimensions, integrated network and layered infrastructure.

2.2.1 State of the Art

There are some existing hardware solutions that offer high level of integration and a quite high energy- and material-efficiency.

Blade-Servers

Vendors like Dell, Hp, IBM, Fujitsu offer highly integrated blade server-systems: Servers are build as slices and are integrated vertical into an blade chassis. The most compact ones (e.g. IBM Blade Center E) are allowing up to 84 server blades in a 42 U 19” rack cabinet. Blade server solutions have switches and network devices integrated, as well as integrated monitoring and controlling solutions. This integrated infrastructure makes blade chassis quite expensive. The mechanical efforts and the needed material for building up the chassis and the blades are quite high.

High density blade-like systems

Some vendors offer high density systems. They mostly use a simpler kind of blade center, which is not as high integrated as the classical blade centers. Supermicro offers a system with 16 low power CPUs (Intel Atom) in 2U, see Tyan (2012). This allows a density of 4 systems per rack rack unit A little more

density is delivered by Tyan: 18 nodes (with low power Intel Xeon quadcore CPUs) in a 4U enclosure, see http://www.tyan.com/solutions/micro_server.aspx. These systems have only a low integration level, usually each single node is completely separated with no common interconnect or monitoring and controlling.

Special High Density solutions

Seamicro, an American start-up, has presented a 10 U system with 64 quadcore low power Intel Xeon servers, see Seamicro (2012). This makes 64 servers per rack unit. This solution has integrated network components and an integrated power management.

HP has presented a non-functional prototype of a very high density system based on ARM-CPUs, see HP (2012). The project with the name Moonshot has 288 ARM-CPU-modules in 7 rack units. Because of the non-x86 architecture of ARM, there is additional work to be done for getting usual server applications running.

The listed examples show that there is a trend in getting IT infrastructure more efficient by searching for ways of higher density and a higher grade of integration of network and management functionalities.

Special High Density Solutions from HPC-Vendors

Especially in HPC-Environments energy- and material-efficiency matters. The new HPC-System at HLRS (Stuttgart) consumes 5 MW (fully deployed, see Resch (2011)) and the planned Super-MUC in Munich 3 MW in the first stage but can be extended to a maximum power rating of 10 MW. If new supercomputers are planned there is often the need to construct a new building for the needed room for the supercomputer and the needed additional infrastructure.

Today HPC is build up mostly with standard components: There are many existing HPC-systems that are based on blade servers (e.g. the “Mare Nostrum” supercomputer of the Barcelona Supercomputing Center, see BSC (2012)). Some still use systems with even lower density, for achieving lower capital expenditure or to realize configurations that are not doable with blade servers.

Some vendor-specific systems are non-standard solutions with new form factors and special components that are allowing a higher density. A well known example for this is the IBM Blue Gene with 1024 nodes in one rack, see IBM (2011).

2.2.2 Dimensions: Standard, non-standard

The dimensions of the rack-level compute box will be similar to a typical 19” rack which comes in different standardised sizes. Common widths of racks are 60, 70 or 80 cm, common overall lengths are 60, 80, 100 or 120 cm. The height is usually 2 meters which results in net 42 or up to 48 rack units (U), where 1 U has a height of 4.445 cm.

To reach a high density per rack, the length of the compute shelf will be at least

100 cm, therefore we will need 120 cm length of the rack. To be able to integrate side-blowing fans or similar cooling techniques that might need air baffles, 80 cm width seem to be most suitable. The typical height of 2 meters will also be adopted from the standard racks.

The typical 19” mounting system will at least partially be adopted because it enables us to easily integrate standard server and infrastructure components into the rack. We will evaluate whether to create an own mounting standard for the very long compute shelves. Because these shelves will be populated with compute nodes on their full length, it is for maintenance purposes necessary to be able pulling them out up to the stop to reach the rearmost compute modules.

2.2.3 Integrated network

The rack-level compute box will have an integrated network for monitoring, management and controlling of all components. The network will be a multi-layer architecture as follows:

- The RECS compute shelves, which provide the main computing power inside the compute box, have an integrated microcontroller based monitoring and controlling solution. This microcontroller solution will be the first monitoring level. To the rest of the network it can be connected via Ethernet, so Ethernet will be the second level.
- Infrastructure servers, as well as storage servers will be monitored via their onboard Ethernet network cards and monitoring software on operation system level, also on the second level.
- There will be a central monitoring node which collects all information and sensor data from all nodes and sensors inside the rack-level compute box. Depending on the specific sensors, this will happen on Ethernet level and thus on the second network layer.
- All Servers will be connected to the storage for data access. This will be the third network layer.
- All Servers will be connected via a fast interconnect solution for synchronising computation jobs. We are evaluating which techniques are suitable for the planned scenarios. This will be the fourth and fastest network layer.

From the above described layers, it is possible to merge the second, third an fourth layer to a generic network layer which could be Gigabit Ethernet. Whether these layers will be merged or there will be separate layers has to be evaluated.

2.2.4 Layered Infrastructure

The infrastructure of the rack-level compute box will consist of several layers. As

a basis we need the rack itself which will be stackable in a row. In this rack we will integrate a cooling solution which is highly efficient and builds an ideal basis for later heat re-use in the container-sized compute box (ComputeBox2).

The components inside of the rack-level compute box will be layered as well. Each compute shelf can integrate different types of compute nodes of the COM Express Computer-On-Module standard, see the website www.comexpress-pnp.org/com_express.html for a description. The Com Express modules can be re-configured just as needed because of a simple plug-and-play concept.

The compute modules will come without local storage, thus there will be a storage server that also acts as a boot server. This boot server is a basis for the compute modules to boot from and thus is a layer underneath the compute modules.

In some use cases we will use a virtualisation infrastructure to encapsulate the whole hardware and work only on the virtualised machines. Through this additional layer we can gain more flexibility in managing resources which hopefully will lead to more energy efficiency.

2.3 Integrated Cooling

IT equipment releases as much heat as the power it absorbs during its operation. To assure the integrity of IT equipment and to provide users with a reliable service, such heat has to be continuously removed by means of cooling systems. The recommended design conditions for IT room cooling are different than for human occupancy. Hotter temperatures are usually admitted in comparison with cooling for human comfort. As consequence apposite strategies and components have to be installed for dehumidification, which is necessary to avoid static electricity in the data center, even if to a smaller extent compared with comfort cooling ((DELL.COM/PowerSolutions, 2009)). Required temperature and humidity conditions can change along manufacturers. Nevertheless the range of variation around the set points is very restricting in comparison with cooling for human comfort.

The A.S.H.R.A.E. (American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.) has recently updated the recommended and allowable thresholds of operation in Data Processing Environments according the following figure (ASHRAE Journal, December 2011),

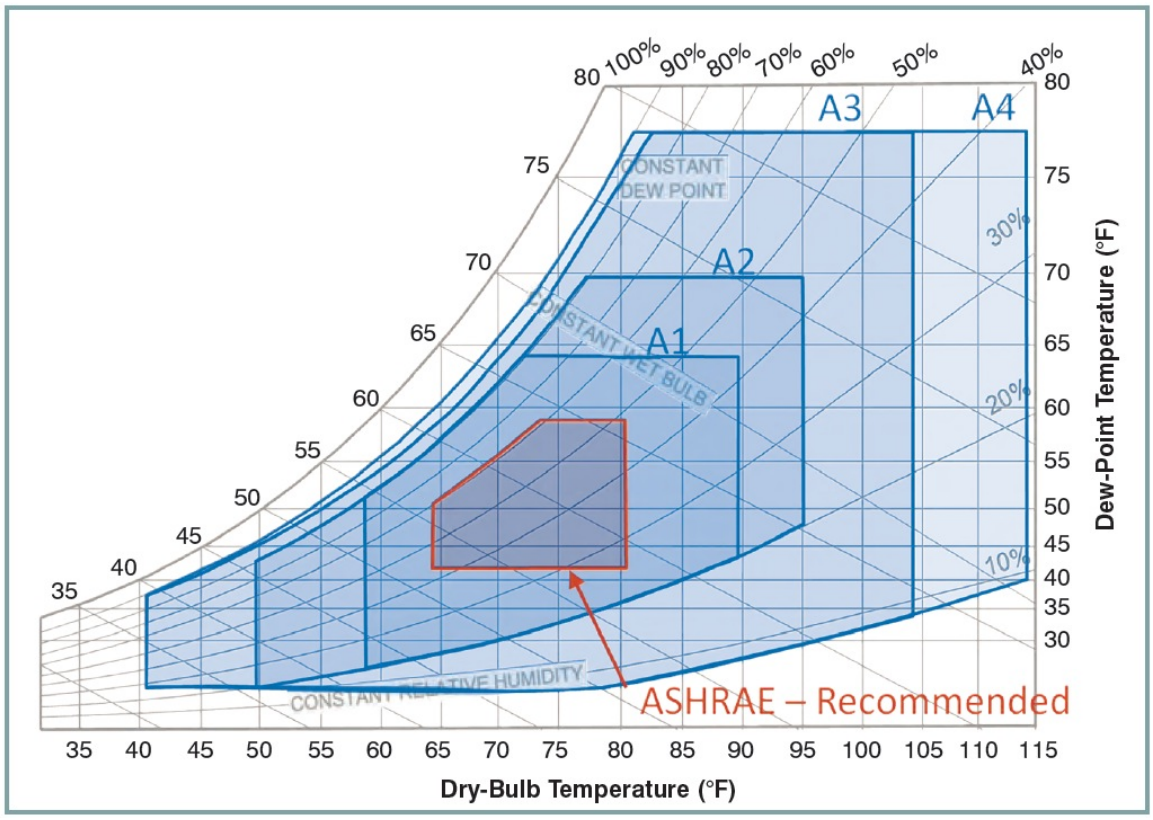

Figure 1. 2011 ASHRAE environmental classes for data center applications. Note: Envelopes represent conditions at IT Equipment inlet.

For this reason, IT room cooling is usually referred to as precision cooling as well, meaning that a strict control of the indoor air conditions is required. Last but not least, as IT equipment operation occurs along the overall year, cooling is always required, resulting in 8760 working hours, much higher than the working hours in the cooling mode in a south Europe office buildings. The large amount of heat to be dissipated and the long operation time are one of the main causes of high energy consumption in data centres (DC).

Thereby cooling is one of the most important topics in a DC, being essential for the correct operation of the IT equipment, requiring strict control, running all over the year and causing high energy consumption. For this reasons the CoolEmAll project will investigate efficient cooling solutions for optimal design of DC.

In this chapter, existing cooling options at rack level are reviewed in order to select energy efficient solutions suitable to the development of ComputeBox1. Actually centralized cooling solutions are usually applied in a DC, meaning that rack cooling is dependent on the overall computer room cooling system. Thereby this chapter lists centralised cooling solutions as well, showing the effect in the surrounding of the rack, especially with respect to the air management. The scope is to select cooling solutions to be implemented in ComputeBox1.

2.3.1 Cooling practices in Data Centres

The traditional approach for cooling in data centers is based on air cooling. Due to high cooling requirements, low specific heat capacity of the air compared with other fluids and ventilation inefficiencies, high ventilation rates are needed with a strict control of temperature, humidity and flow direction. Such control gets stricter and stricter with the increase of the DC density (higher than 20 kW/rack). In high density DC liquid cooling solution is being currently widely applied. It has to be highlighted that the terms “air” or “liquid” cooling refer to the means which are used to directly cool the racks. In fact an air cooling approach is usually supported by liquid cooling, as the air is chilled by cold water or another refrigerant.

Figure 2 gives a general view of a cooling system in a DC. The cooling system can be divided into three main parts that can be located in different places of the building:

- Cold Water production and distribution, usually situated outside the data center (typically on the building roof or in the basement);
- Cold air production situated in dedicated technical room outside the IT Room or in the same IT Room;
- Cold air distribution situated in the IT Room.

When talking about cooling solutions at rack level, the focus is on the air flow management around and within a rack and liquid cooling solutions.

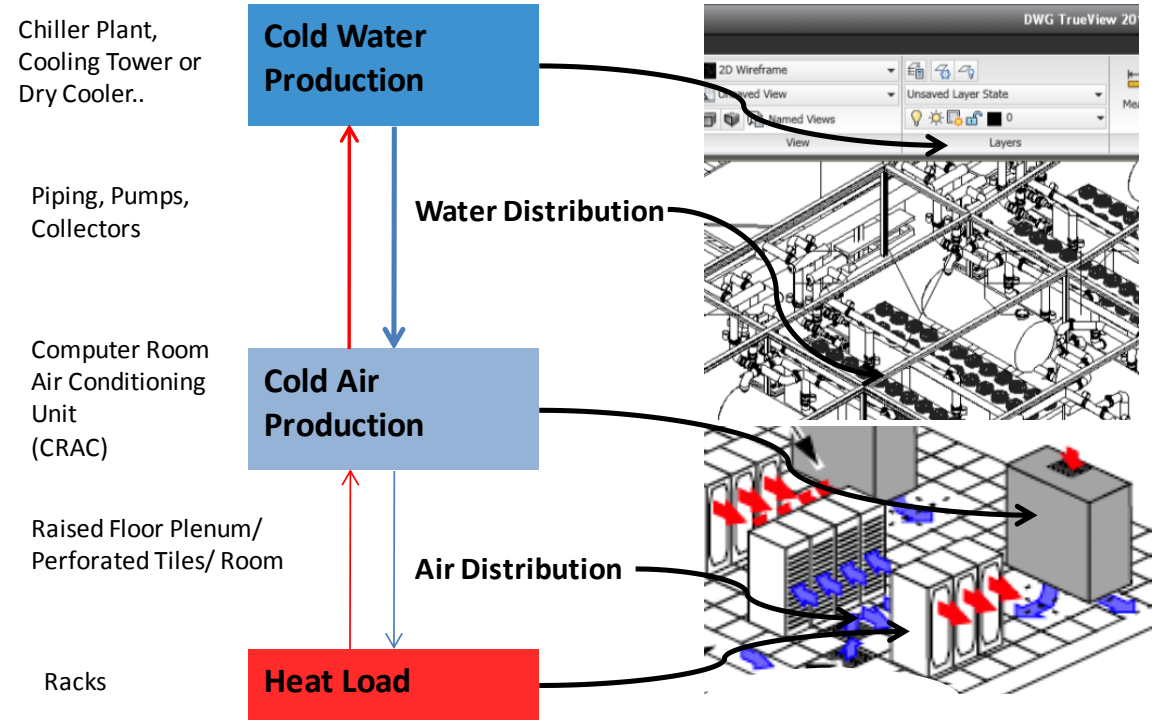

Figure 2. Typical cooling concept and lay out in a DC

2.3.1.1 Air management solutions

Ventilation in open racks (no or perforated doors) is determined by the air distribution strategy adopted in the overall computer room.

(APC, 2003) summarizes (Figure 3) nine combinations of flooded, partially ducted and fully ducted distribution lay outs. Given that recommended air flow directions in a rack are front to rear, front to top, or front plus top to rear, impulsion of air to the racks usually occurs via raised floor: cold air flows under the floor and reach the racks through perforated tiles. Expulsed hot air returns to the Computer Room Air Conditioner (CRAC) to be cooled and resent to the racks, like in Figure 4, with the creation of “ideal” cold and hot aisles. Return hot air can also be ducted via empty ceiling as in Figure 5.

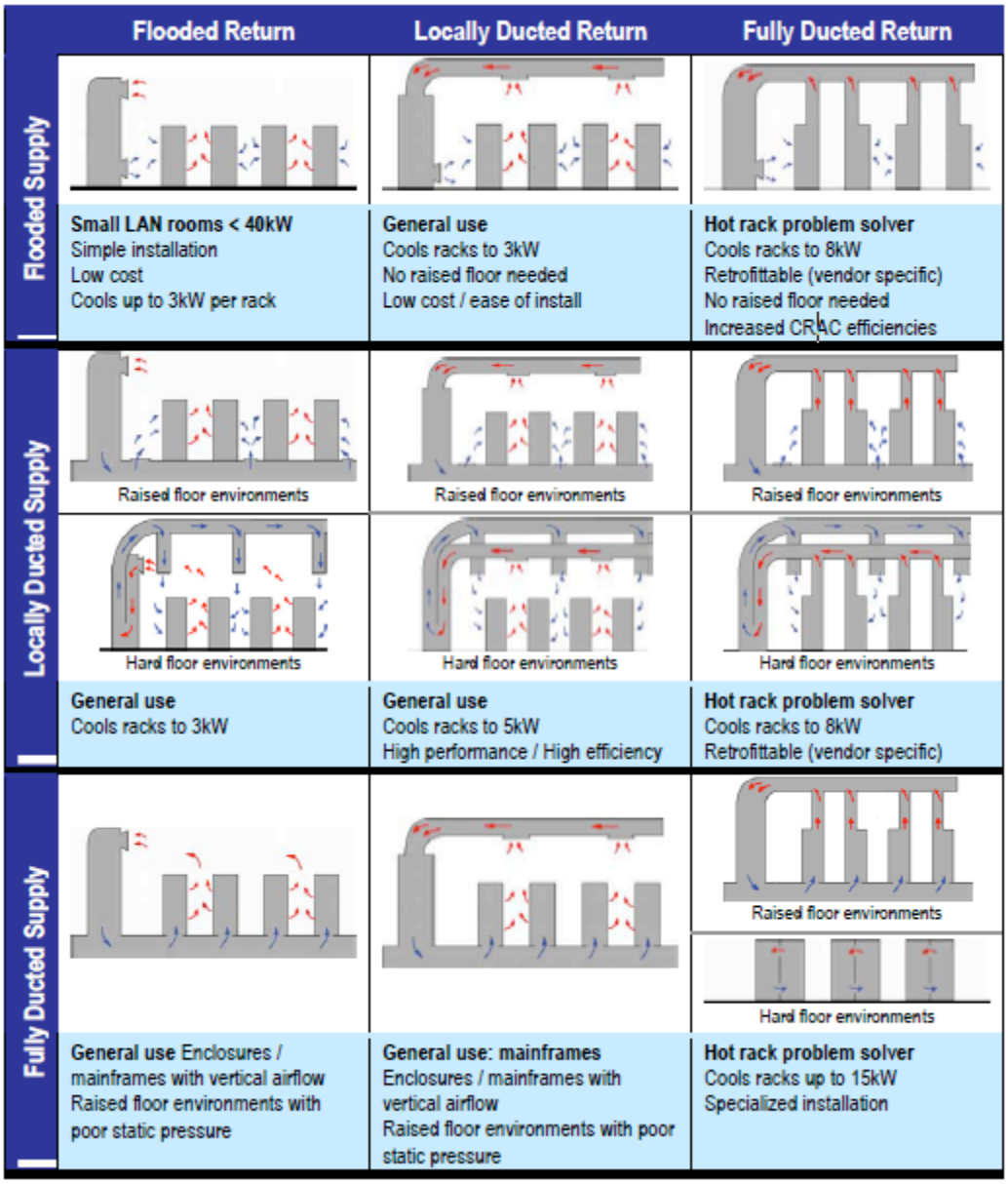

Figure 3. Layout for air distribution in DC

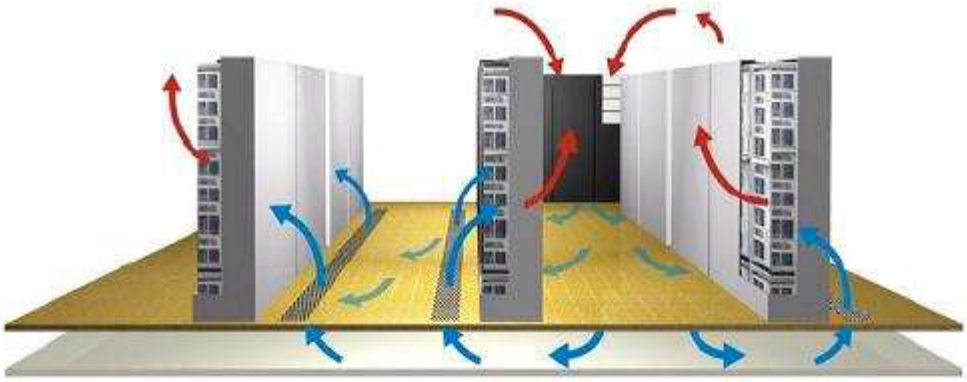

Figure 4. Raised plenum floor and air flow in an IT room with CRAC

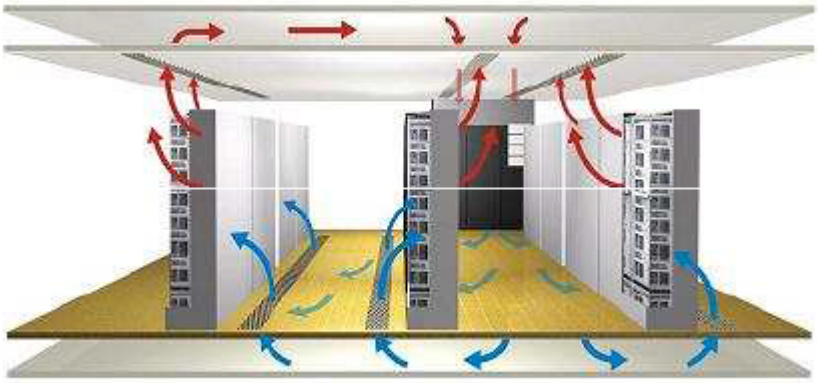

Figure 5. Raised plenum floor and empty ceiling in an IT room with CRAC

Ducting the airflow diminishes the risk of air mixing.

Air mixing is one of the major problems in air management in DC to be avoided. It can occur in several locations in the IT room:

- Cold airflow occurs from the room to the raised floor due to pressure gaps: the phenomenon is called “negative pressure flow” and increases, in a negligible way, the temperature of the air addressed to the IT equipment.
- Around the CRAC unit, cold supply air returns directly to the CRAC, bypassing the IT equipment, like in Figure 6: such phenomenon, referred to as airflow bypass, significantly lowers the volume flow reaching the IT equipment and the return air temperature to the CRAC. Consequently the efficiency of the CRAC unit and the supporting cooling system is decreased as well, and the energy use increases.
- Heat diffusion occurs from the hot aisle to the cold aisle as in Figure 8. The phenomenon, known as recirculation airflow, increases the temperature at the inlet of electronic equipment that is located in the top of

the racks and at the sides of an aisle (hotspots), as a result of diffusion of hot air. The risk is the overheating of such equipment.

- Air recirculation can occur within a rack as well due to local fans or to free space.

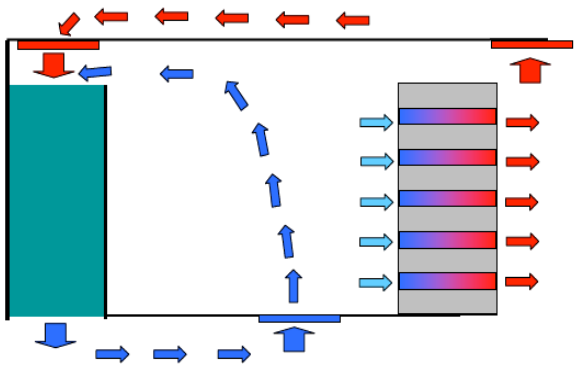

Figure 6. Bypass airflow phenomenon (Tozer, Gestión del Aire en Centros de Cómputos - Principios, 2009)

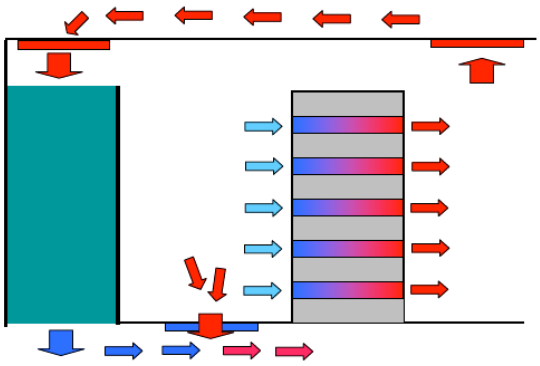

Figure 7. Negative pressure flow phenomenon (Tozer, Gestión del Aire en Centros de Cómputos - Principios, 2009)

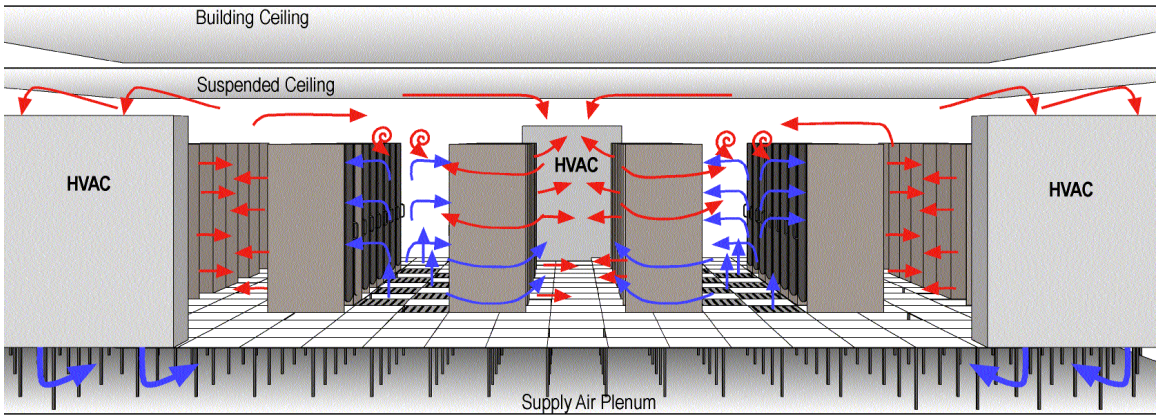

Figure 8. Recirculation problems around the racks in IT room with cold air impulsion via raised plenum floor and hot air expulsion via suspended ceiling (Kennedy)

Thereby in general, mixing air flow leads to IT equipment over heating, with consequent fails, and to higher energy use for cooling. For this reason it is recommended to concretely separate the cold air impulsion path and the hot air expulsion path.

On this subject, the creation of containment of hot and/or cold aisles (Figure 9), where aisle between rows of racks is bounded with exclusively hot-air outlets or

exclusively cool-air intakes, turns out to be one of the best solutions to avoid air mixing problems, recirculation in particular. Isolation is obtained by the installation of plastic, glass or hard wall separations between the aisles.

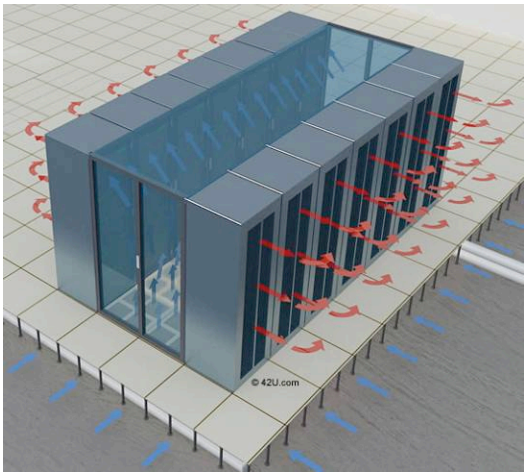

Figure 9. Contained cold aisle

Recirculation and bypass risk can be reduced by a correct location of the CRAC unit. When the CRAC units are installed in the IT room, it is recommended to install them perpendicularly to hot isles to avoid air short circuit from the closest perforated tiles. Nevertheless, the optimal solution would be to install the CRAC units in a dedicated technical room like in Figure 10, as the bypass airflow is strongly cut out and ventilation is in general improved because privileged flows directions are established

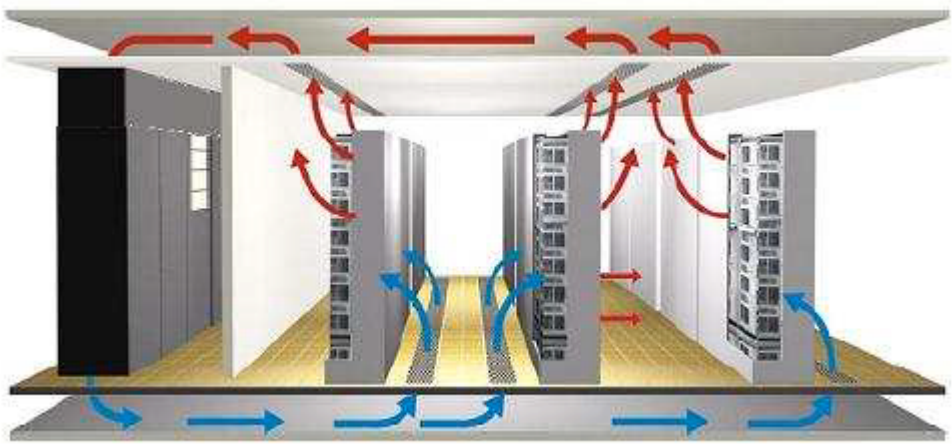

Figure 10. Cold air impulsion via raised plenum floor and hot air expulsion via empty ceiling, with CRAC unit located outside the IT room

Within the rack itself the possibility exists for hot exhaust air to be recycled into the equipment air intake, especially when vertical space is not occupied by any component. Covering these gaps with blanking panels helps to maintain proper airflow (Figure 11). Rack cabinets fans can also be used to increase the heat removal from IT equipment but they can cause recirculation problem as well, especially when coupled with the computer room air handling unit (CRAH).

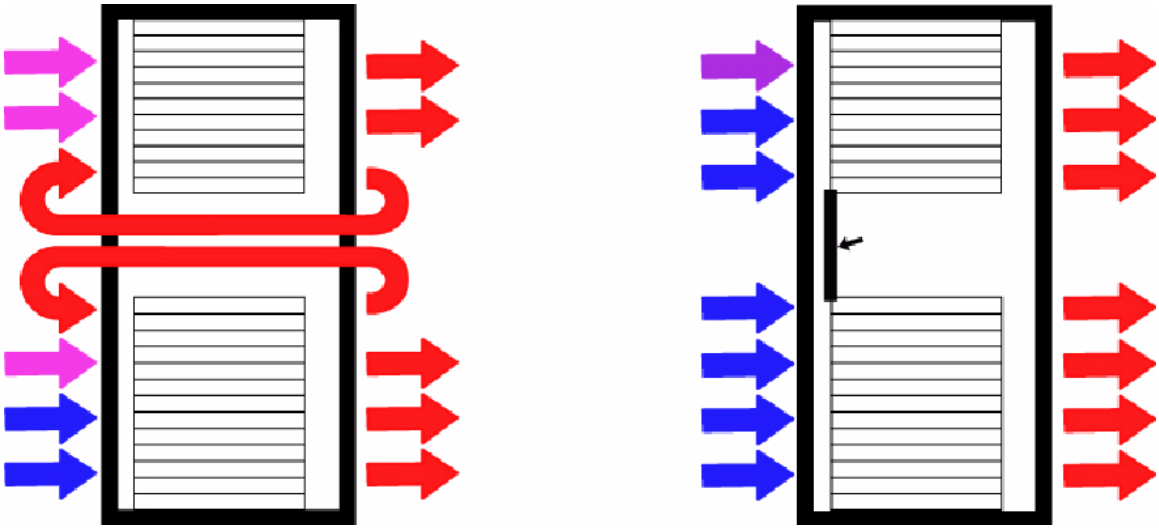

Figure 11. ***Diagrams of rack airflow showing effect of blanking panels*** **(APC, 2005)**

New generation cooling solutions, usually referred to as close coupled cooling solutions, include in-line and in-row air conditioners to improve the impulsion and expulsion of air. Such solutions exist in different configurations but the common approach consists of bringing the functionality of a CRAC close to the racks, as in Figure 12. Moving the air conditioner closer to the equipment rack ensures a more precise delivery of cold air where it is needed and a more immediate capture of exhaust air, reducing the risk of recirculation.

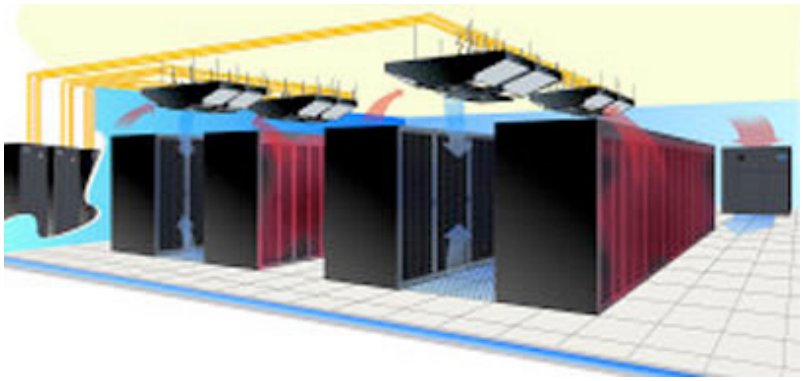

Figure 12. In row overhead air cooling

2.3.1.2 Liquid cooling solutions

Liquid cooling is defined as “the case where liquid must be circulated to and from the entity for operation”. For instance, a liquid cooled rack defines the case where liquid must be circulated to and from the rack for operation. This definition can be expanded to liquid cooled IT equipment and liquid cooled electronics. The overall goals of the liquid implementations are to transfer as much waste heat to the facility water as possible and, in some of the implementations, to reduce the overall volume of airflow needed by the racks. In addition, implementation of liquid cooling may be required to achieve higher performance of the IT equipment through lower temperatures achieved with the cooling of microprocessors.

Liquid cooling solutions can be divided into two categories: open loop vs. closed loop.

Open-Loop solutions bring the heat transfer closer to the equipment rack but are not completely independent of the room in which they're installed. The air streams will interact to some extent with the ambient room environment. These products will use either chilled water or refrigerant through their cooling coils. All will require remote heat rejection via a mechanical chiller system, condensing either chilled water or refrigerant.

One example is the rack with a Rear Door Heat Exchanger (RDHx) (Figure 13). A RDHx is placed in the airflow outlet of a server rack. Hot server-rack airflow is forced through the RDHx device by the server fans. Heat is exchanged from the hot air to circulating water from a chiller or cooling tower. Thus, server-rack outlet air temperature is reduced before it is discharged into the data center. These units can remedy hot spots in existing data centers, supplementing the existing air conditioning or for smaller loads and rooms, provide cooling for spaces not originally designed as data centers- data rooms, closets, labs. Installed on racks, these units do not take up floor space-an important point in small size installations.

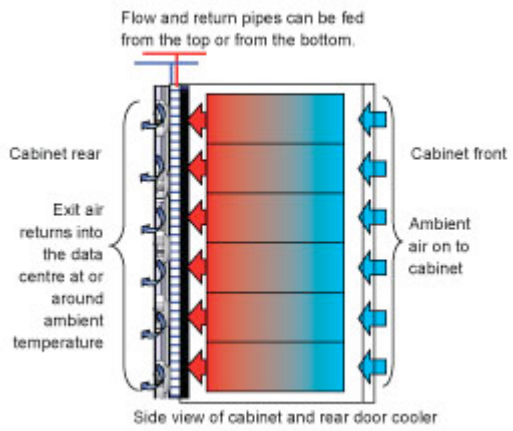

Figure 13. Rear door heat exchanger cabinet

Closed-loop cooling addresses the compute load independent of the room in which it's installed. The rack and heat exchanger work exclusively with one another, creating a microclimate within the enclosure. For this reason such solutions are also referred to as in-rack solutions.

In in-row liquid cooling solutions (Figure 14), the AC is adjoined to the server rack and both are fully sealed. The solid doors on the enclosure and in-row air conditioning device contain the airflow, directing cold air to the server inlet and exhaust air, via fans, through the cooling coil. The close-loop design allows for very focused cooling at the rack level. Users can, therefore, install very dense equipment exclusive of the ambient environment. As a result, they have the flexibility to use unconventional rooms and spaces to house the IT equipment.

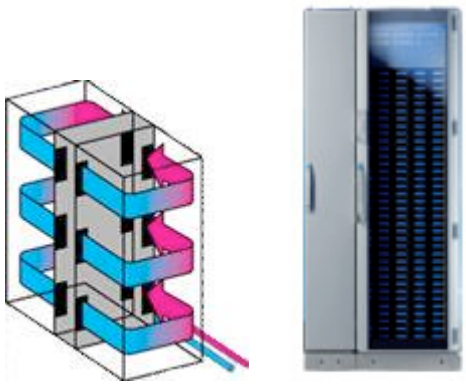

Figure 14. An example of in-row air conditioning cooling cabinet

It must be taken in account that all these strategies to avoid air management problems can be combined in the same facility.

2.3.2 Possible Cooling Solutions for ComputeBox1

The air management solutions described above are mostly centralized systems than rack integrated systems, but they determine the heat removal from the rack and influence the IT equipment performance. Thereby, air management problems in the overall IT room have to be considered when looking for a cooling solution at rack level. (Tozer, Kurkjian, & Salim, 2009) points out that negative pressure ratios (NP), bypass air flow (BP) and recirculation air flow (R) largely affect the heat removal efficiency of ventilation, as previously commented. In fact, because of these phenomena, the temperature reaching the servers is higher than the temperature supplied by the CRAC/CRAH units. Because of that, on one hand the risk of IT equipment overheating arises; on the other hand, given the temperature for the cold air required by the manufacturers, the cooling system has to supply as colder temperature as higher bypass and recirculation air flows occur, leading to higher energy consumption. It is thereby necessary to reduce such flows. (Tozer, Kurkjian, & Salim, 2009) reports some metrics to assess the negative pressure ratios, the bypass air and the recirculation air, deriving the following relationship:

$$BAL(1+NP) = \frac{1-R}{1-BP}$$

with:

$$BAL = \frac{\Delta T \ on \ the \ racks}{\Delta T \ on \ the \ CRAC}$$

Representative indices were obtained in the cited study for typical air management solutions (Figure 15).

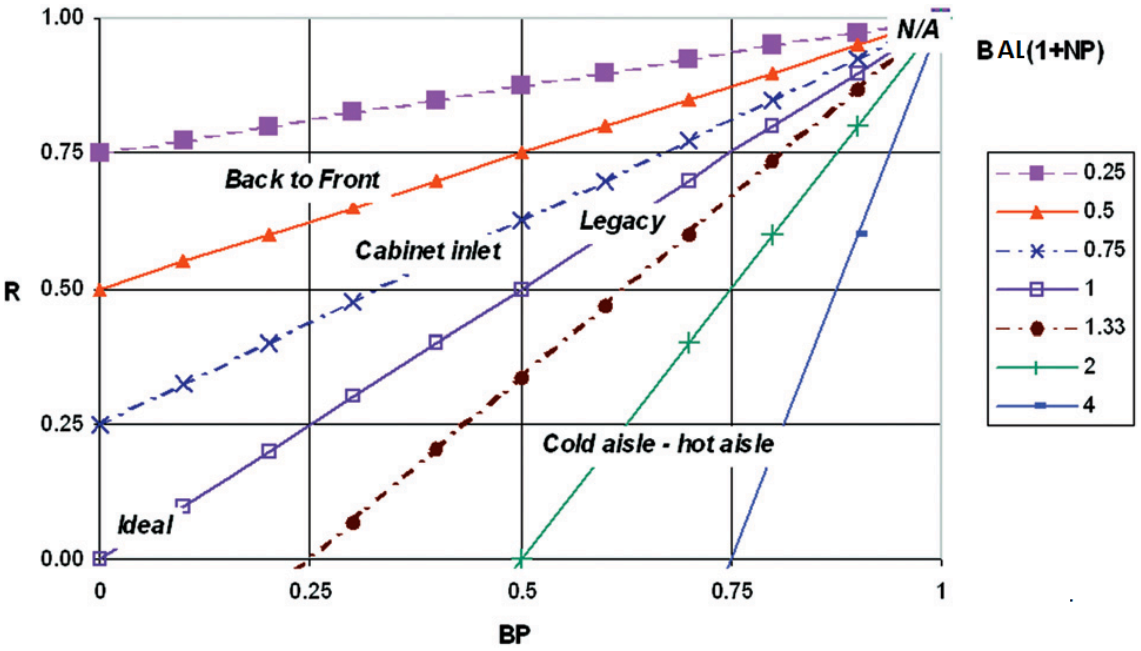

Figure 15. Air management metrics for different air management solutions (Tozer, Kurkjian, & Salim, Air Management Metrics in Data Centers, 2009)

The ideal solution is obtained for BP=0, R=0, NP=0, thereby BAL=1. According to (Tozer, Kurkjian, & Salim, 2009) bypass airflow is minimized in cabinet inlet solutions, because the cold air enters from below the cabinet directly. Nevertheless, as in this case, blanking panels are often missing, recirculation within the rack can be high leading to the creation of hotspots. In back to front air flow, recirculation around the racks is the main problem. The arrangement of racks with the creation of hot/cold aisles prevents the hot air to diffuse in the cold one but recirculation still occurs. To diminish it, an increase of cold air volume flow could help. Nevertheless this would increase the energy consumption on one hand and on the other hand the bypass airflow could be increased because of the extra air supply (Khattar, 2010).

The containment of the aisles can respond to both the recirculation and bypass airflow problems, but according to (De Coufle, 2009) cold/hot aisles are the optimal solution to solve air mixing problem but they are not largely applied because of costs, reduced accessibility to the racks and the risk and possible problems for the rack fans because of the different pressures between the two zones. Usually, contained cold aisles are preferred to contained hot aisles: as raised plenum floor are often implemented, the path for the cold air is already better identified than the hot air path. Containments can guarantee the necessary heat removal in high density IT room as well (RITTAL). On the other hand, in row liquid cooling solutions is advisable in high density IT room, being the precise cooling solution for excellence. Its implementation is expected to efficiently remove heat from servers, reducing consumption for ventilation and avoiding air mixing problems in the IT room that can still occur in RDHx devices, as the latter are open. On the other hand RDHx are flexible devices that can be installed in

combination with CRAC or alone, in new or existing DC, but not for all rack designs (US Department of Energy). More efficient of CRACs, RDHx may make creating “hot aisle isolation” less important because they can sufficiently reduce server outlet temperatures.

A review of cooling systems has been carried out in this chapter and some technologies have resulted to be promising with respect to energy efficiency. Please note, cooling concepts have been shown rather than existing products as they are several and differ manufacturer by manufacturer. Furthermore, it is here highlighted that the real efficiency of a cooling solution also depends on the design (e.g. open surface of the tiles), the installation (e.g. presence of air gaps or obstructions in the raised floor), the setting of working parameters (e.g. cold air temperature, ventilation rates) and control strategies (e.g. variation of fans frequency). Thereby the promising efficient solution could result inefficient if not well designed, installed and managed. Such issues will be tackled in the CoolEmAll project, as its main scope is to provide a tool for optimal design of energy efficient DC.

2.4 Components within the rack-level compute box

In this section we describe components of the rack-level compute box, providing overview on rack-level components, server and storage solutions, and infrastructure required to operate solutions energy efficiently.

2.4.1 State of the Art

State of the Art of a full-featured rack with all typical integrated components contains the following:

- Uninterruptable Power Supply (UPS), if this is not centralised
- Storage for active data which is commonly used by all compute resources inside the rack
- Backup storage which is again commonly used
- Compute nodes. These can be of various types, from single standard 19” servers to Blade Centres
- Keyboard-Video-Mouse module for local administration
- Interconnect, possibly various interconnects for different purposes, see section 2.2.3

In large environments these components to build up a data centre environment are always found. Sometimes they are centralised, sometimes they are distributed over several racks – both approaches have advantages and disadvantages. Our goal is to propose a one-rack solution that can be scaled up to hundreds racks by adding more racks that all have their own infrastructure.

2.4.2 Server solutions

For an overview about today's high-end and big server systems and their physical dimensions, please see section 2.2.1. For this section we will give a more internal overview about the available computing solutions.

Modern server architectures that are state of the art can be divided into several main lines. Mainframes are very complex computers that have special working environments often used to provide critical business areas. Mainframes retire for the role of standard-based server that can be integrated in our environment due to their very high investment costs and often non-standard functionalities. Common multi-purpose servers are available in various flavours and with differing configurations. Intel's x86 architecture is the most common architecture used nearly everywhere today. POWER is an architecture developed by IBM which still has supporters but is not as popular as the x86 architecture. ARM is an architecture which gains attention in the last months due to very low power consumption at reasonable computation power. As already mentioned, HP is building an ARM based high-density server with multiple ARM processors inside.

Most of these architectures come in different power classes. Usually the low-end server class has one CPU that mostly has several cores per CPU integrated. Above this class there is a dual-CPU server class that combines the power of two CPUs on one mainboard which leads consequently makes much memory available within one server. Multiple-CPU servers are also available on the market, but are usually designed to reach a maximum memory capacity and a maximum number of CPUs on one mainboard. Besides this goal the energy-efficiency of these servers is mostly by far worse than the energy-efficiency of one- or two-CPU servers.

To reach a good trade-off between flexibility, energy-efficiency, density and sizing of the servers, we plan to use multiple RECS servers as the central compute servers, see Christmann (2009) and section 2.5.1. This allows us to test several CPU architectures, CPU power classes and memory capacities under the same working environment and with the same monitoring and controlling infrastructure.

2.4.3 Storage solutions

The focus of this project is on energy-efficient server solutions. Therefore storage shall be regarded only as a given infrastructural component that can be used. For energy-efficient management-approaches see for example the GAMES project (www.green-datacenters.eu).

The storage solution we plan to integrate into our testbeds of the rack-level compute box will have the same network connection like the compute nodes and will also work as a boot server for disk-less remote boot.

2.4.4 Infrastructure

Infrastructure means all things that usually are not mentioned when describing a server- or storage-project but which are essential and can enable the user to run the servers energy-efficient or not.

For the interconnect see section 2.2.3. As described there, we will have several interconnection layers: microcontrollers to get monitoring information from the compute nodes and to be able to control them remotely and Ethernet for a commonly used data interconnection and for monitoring the application based metrics that can only be accessed via the operation system. Furthermore we will possibly have a dedicated storage network and a high-speed interconnection for the compute nodes like Infiniband, PCIe-Switching, 10 Gigabit Ethernet or Numascale. Which techniques we will use has to be evaluated.

The cooling infrastructure is described in section 2.3. The general idea is that the cooling is integrated into the rack and can be monitored and optimised to meet the actual cooling needs.

The power distribution is another infrastructural part to be mentioned. The use of a UPS will be ignored here because it is highly likely that our testbeds have own UPS already installed. In general we want to highlight that for high energy-efficiency, a correctly dimensioned Uninterruptable Power Supply (UPS) is important. Modern large-scale UPS work at an efficiency of up to 99%, whereas smaller systems reach about 90%. Because the efficiency depends on the actual workload, UPS should always be operated in the upper third of their maximum capacity, because this is where their efficiency reaches its maximum level. Furthermore, Rasmussen (2011) highlighted that some UPS have special energy-efficiency modes.

In the GAMES project we integrated intelligent multi-outlet power strips that had power sensors integrated to measure not only the power consumption of the RECS but also of the remaining infrastructure and the servers that had no power meter integrated. This is a good possibility to acquire knowledge about the actual power consumption if the hardware does not support such features.

2.5 Integrated monitoring and management

Energy efficient operation of servers requires infrastructure allowing monitoring, controlling and managing cluster servers energy efficiently, adapting to fluctuating resource-demands, applications and environmental conditions. In this section we describe RECS servers with integrated monitoring and controlling capabilities, and provide an architecture allowing to monitoring and manage rack consisting of several RECS and other components.

2.5.1 Cluster Server RECS

Before describing the monitoring and controlling infrastructure that comes with

the RECS cluster server, we will give an overview about the overall system architecture. The following ideas and descriptions are based on collaborative efforts of the GAMES project, the authors are also part of it.

The cluster server RECS consists of 18 single CPU modules, each of them can be treated as an individual PC. The mainboards are COM Express based CPU modules, each mounted on a standardized baseboard which makes it possible to use every available COM Express mainboard that has the “basic” size. In CoolEmAll we will evaluate which CPU module will be the best for each particular use-case. Each baseboard is connected to a central backplane. This backplane has two functions, first it forwards each Gigabit Ethernet Network of the CPU modules to the front panel of the server, and second it connects the baseboards’ microcontrollers to the central master-microcontroller. For debugging purposes it has been quite useful in the past to have direct access to single mainboards, therefore every baseboard has several connectors as listed in the following table.

Connector / Button	Position
USB	On each Baseboard & two at the Front Panel of the Server Enclosure (for Compute Node 9)
2x SATA	On each Baseboard
VGA	On each Baseboard & one at the Front Panel of the Server Enclosure (for Compute Node 9)
18x Gigabit Ethernet	Front Panel of the Server Enclosure
Fast Ethernet for the Monitoring Solution	Front Panel of the Server Enclosure
Power Connector 12V	Back Side of the Server Enclosure & on each Baseboard
Power & Reset Button	On each Baseboard
Control Buttons for the Monitoring Solution	Front Panel of the Server Enclosure
LCD Display for the Monitoring Solution	Front Panel of the Server Enclosure

Table 1. Physical Interfaces of the RECS Cluster System

All components within the cluster server share a common Power Supply Unit (PSU) that provides 12V with a typical efficiency of more than 92%. The several potentials needed for the mainboard chipset, CPUs and other components are provided by both, the baseboards and the mainboard potential transformers in the cluster server itself.

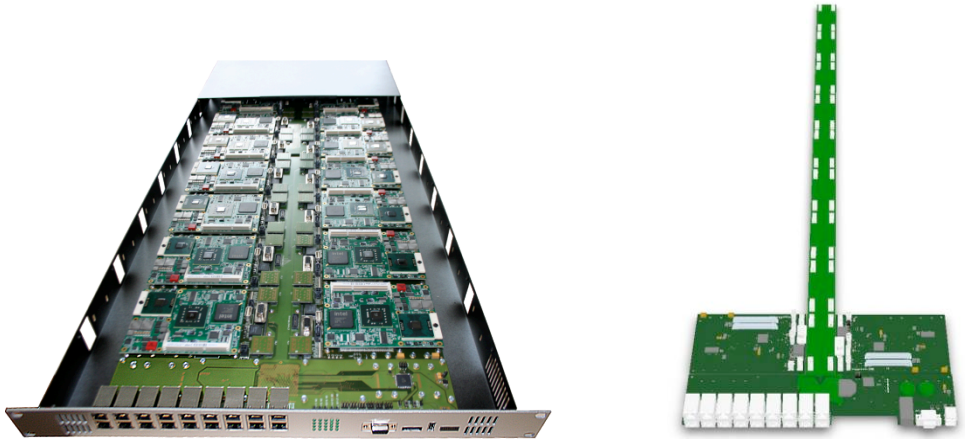

Figure 16. Picture and a technical sketch of an early prototype of the Cluster System, the Cooling Units and the LCD display is omitted

The novel monitoring approach of the RECS Cluster System is to reduce network load, avoid the dependency of polling every single compute node at operation system layer and build up a basis on which new monitoring- and controlling-concepts can be developed. Therefore the status of each compute node of the RECS Cluster Server is connected to an additional independent microcontroller in order to manage the measured data. The main advantage of the RECS Cluster System is to avoid the potential overheads caused by measuring and transferring data, which would consume lots of computing capabilities; in particular in a large-scale environment this approach can play a significant role. On the other hand, the microcontrollers also consume additional energy. Comparing with the potential saved energy, it is expected that the additional energy consumption could be neglected. This microcontroller-based monitoring architecture is accessible to the user by a dedicated network port and has to be read out only once to get all information about the installed computing nodes. If a user monitors e.g. 10 metrics on all 18 nodes, he would have to perform 180 pulls which can now be reduced to only one. This example shows the immense capabilities of a dedicated, aggregating monitoring architecture.

The monitoring architecture is realized by a master-slave microcontroller architecture which collects data from connected sensors and reads out the

information every mainboard provides via SMBus or I^2C. Each baseboard is equipped with a thermal and current sensor. A list of measureable data is given in CoolEmAll Deliverable 4.1. All sensor data are read out by one microcontroller per baseboard which acts as a slave and thus waits to be pulled by the master microcontroller. The master microcontroller and thus the monitoring- and controlling-architecture, are accessible to the user by a dedicated network port and additionally by a LCD display at the front of the server enclosure.

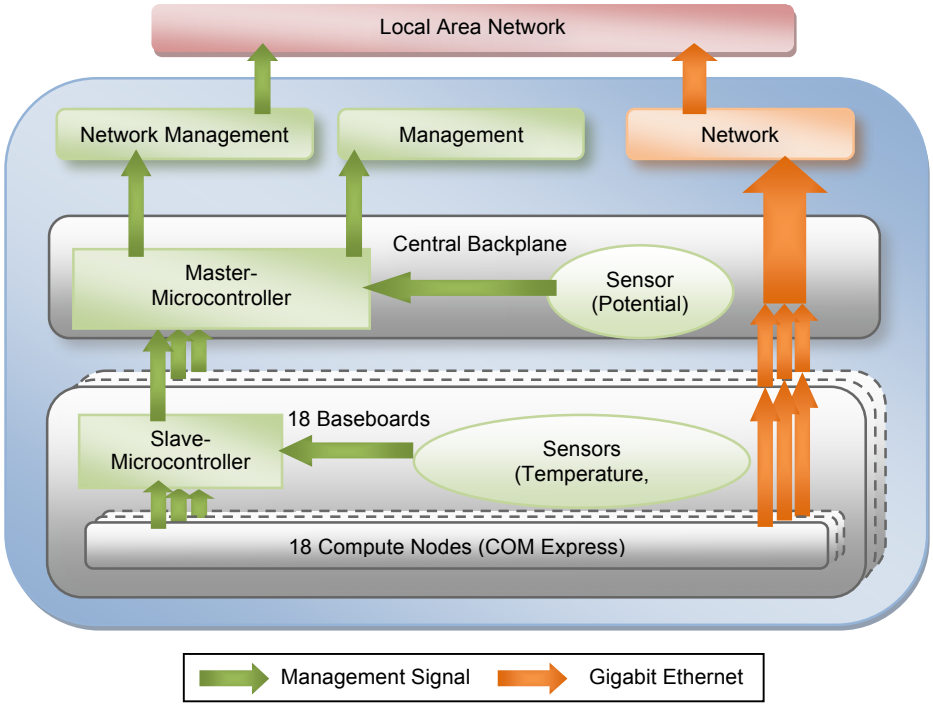

Figure 17. Architecture of the Master-Slave Microcontroller Monitoring System

Additionally to the monitoring approach, the described infrastructure can be used to control every single compute node. Right now it is possible to virtually press the power- and reset-button of each mainboard.

Of course it is even possible to have a mixed setup of energy consumption where some nodes are under full load, others are completely switched off and some nodes are waiting in a low-energy state for computing tasks.

2.5.2 Monitoring and management architecture

The monitoring and management infrastructure of the rack-level compute box is part of the Module Operating Platform, described in CoolEmAll Deliverable D4.1. It is responsible for monitoring, data-history building and controlling of the rack-level compute box. In this section we summarize the monitoring and

management architecture of the rack-level compute box, which is described fully in CoolEmAll Deliverable D4.1.

The architecture of the monitoring and management infrastructure is based on the architecture of TIMaCS framework, as described by Volk (2011). The TIMaCS framework is designed as a policy based monitoring and management framework with an open architecture and a hierarchical structure, operating on top of existing monitoring and management solutions, such as Nagios, Ganglia and others.

Figure 18 shows deployment of TIMaCS within the rack-level compute box. As mentioned earlier, the rack-level compute box contains several RECS, each consisting of compute nodes and a microcontroller with integrated sensors, which provide monitoring information about resources or services running on them. The monitoring tools are collecting data from all RECS microcontrollers and compute nodes within the rack-level compute box. Collected data is transformed into a uniform metric format then, and is aggregated and stored in a round robin database of the TIMaCS node, creating a data history that reflects a time-spatial behavior of captured metrics. Now, stored monitoring data can be retrieved and processed by a GUI, SVD Toolkit, or other external programs, using RPC based API offered by TIMaCS framework. To enable controlling of resources, each managed resource is equipped with Delegates, interfaces allowing executing commands on managed resources. These Delegates can be used to set or change hardware parameters according to predefined policies, workloads- and resource settings. The management capabilities provided by TIMaCS allow analysis of monitoring data, triggering events indicating request for decision. As a result of decision process, commands or actions are selected according to predefined policies or rules, and are submitted to Delegates of compute nodes or resources, where they are executed as a reaction on detected events.

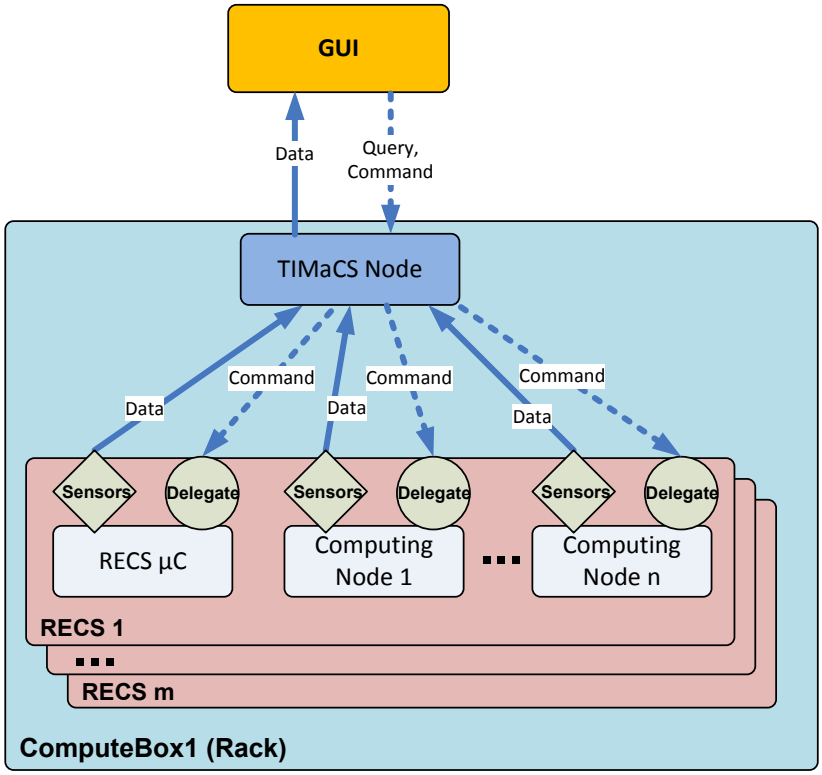

Figure 18. Monitoring and Management of rack-level compute box

The detailed monitoring and management architecture is presented in CoolEmAll Deliverable D4.1.

3 The Data centre Efficiency Building Block (DEBB)

As mentioned, one of the main results of the CoolEmAll project will be the design of diverse types of data-center building blocks on different granularity levels, following a blueprint-specification format called Data center Efficiency Building Block (DEBB). The DEBB is an abstraction for computing and storage hardware and describes energy efficiency of data-center building block on different granularity-levels, as explained in the table below:

Level	Full name	Description
1	Node Unit	This level reflects the finest granularity of building blocks to be modeled within CoolEmAll. This smallest unit reflects a single blade CPU module, a so-called "pizza box", or a RECS CPU module

2	Node Group	This level reflects an assembled unit of building blocks of level 1, e.g. a complete blade center or a complete RECS unit (currently consisting of 18 node-units)
3	ComputeBox1	This level reflects a typical rack within an IT service centre
4	ComputeBox2	Building blocks of this level are assembled of units of level 3, e.g. reflecting a container filled with racks or even complete compute rooms.

Table 2. DEBB Definition

A DEBB of each granularity level is described by:

a) *composition*: specification of components and sub-building blocks

b) *physical dimensions:* outer physical dimensions (=black-box description), and optionally arrangements of components and sub-building blocks within particular DEBB (=white-box description)

c) *power consumption* for different load-levels (e.g. 20% rise per step), addressing various types of components within the same building block

d) thermodynamic *modelling profile*, describing air-flow (including direction and intensity) and temperature on inlets and outlets for different load-levels

e) *metrics* assessing energy efficiency of particular DEBB

The following subsections provide a detailed explanation on these specification-items, used to describe a DEBB.

3.1 Composition

A DEBB is described by its components and sub-building blocks. The following paragraphs describe composition of a DEBB for different granularity levels:

Node Unit consisting of the technical description and corresponding numbers of components within particular node-unit:

- Main-Boards, including network
- Potentially additional cards put in slot of the MB
- CPUs, including possible operating frequencies
- Memory Modules
- Cooling elements, including fan

- Optionally, storage elements, if included

Node Group consisting of the technical description and corresponding numbers of components and Node-Units within particular Node-Group:

- Node-Unit descriptions or a reference to Node-Unit descriptions
- Main-Pane to which all Node-Units are connected

ComputeBox1 (rack) consisting of the technical description and corresponding numbers of components and Node Groups used within particular ComputeBox1:

- Node-Group description or reference to a corresponding Node-Group description
- Secondary components, such as interconnect, switches etc, used to interconnect Nodes and/or Node-Groups
- Power-Supply or Uninterruptable Power Supply (UPS) if not centralised, providing power to all Node-Groups
- Integrated cooling devices

ComputeBox2 (data-centre or container) consists of the technical description and corresponding numbers of components and ComputeBox1 used within particular ComputeBox2:

- ComputeBox1 description or reference
- ICT Infrastructure interconnecting ComputeBox1
- Facility cooling devices

The specification of DEBBs provided in this section are used in Section 3.5 to get an overview on possible configurations and reconfigurations of a DEBB, allowing defining configuration space needed for selection or definition of “optimal” DEBB.

3.2 Physical Dimensions

Each building block is described at least by its outer physical dimensions (=black-box description). In addition, a building block can contain also data about the position of its inner components and sub-building blocks (=white-box description). Whereas the white-box description is mandatory for the specification of thermodynamic profiles, it’s optionally for black-box description, used to hide optimal arrangement of components within a building block, protecting business interests.

3.3 Power consumption

A DEBB is described by its power consumption, as a result of different load-levels (20% steps per level), stressing various types of components, affecting at least:

- CPU-Utilization at different operating frequencies
- Memory-Utilization

The following tables provide examples on possible usage levels of CPU and Memory, used to capture power consumption at specified level.

CPU-Usage_Freq_i	Power
0%	p_cpu_0
20%	p_cpu_20
40%	p_cpu_40
60%	p_cpu_60
80%	p_cpu_80
100%	p_cpu_100

Memory-Usage	Power
0%	p_mem_0
20%	p_mem_20
40%	p_mem_40
60%	p_mem_60
80%	p_mem_80
100%	p_mem_100

Table 3. CPU and memory usage levels

Optionally, other loads might be used to stress:

- IO-Utilization
- Network-Utilization at different operating speed

The total power consumption on Node-Level is metered by a power-meter during the stress of:

- All or selected CPUs and Cores within a Node at predefined usage-levels for all the different possible CPU frequencies on that CPUs.
- All or selected memory modules within a Node at predefined usage-levels

The total power consumption on Node-Group-Level is an aggregation of the power consumption of all Nodes of the particular Node-Group, being stressed at predefined usage-level. The usage level of the Node-Group is defined as an aggregation/sum of the usage-level on all nodes divided by the total number of nodes in a group.

The total power consumption on Compute-Box1-Level is an aggregation of the power consumption of all ComputeBox1 and all devices within particular Compute-Box1, being stressed at predefined usage-level. The usage level of the Compute-Box1 is defined as an aggregation/sum of the usage-level on all nodes

within particular ComputeBox1, divided by the total number of nodes in a Compute-Box1.

The total power consumption on Compute-Box2-Level is an aggregation of the power consumption of all ComputeBox1 and all devices and components within particular Compute-Box2, being stressed at predefined usage-level. The usage level of the Compute-Box2 is defined as an aggregation or weighted sum of the usage-level on all nodes within particular ComputeBox2, divided by the total number of nodes in a Compute-Box2.

The characterization of power consumption on each level provides an instrument to assess the power consumption of an application on particular building block, once the resources usage level of an application is known.

3.4 DEBB for thermodynamic modelling

The DEBB will be modelled in CoolEmAll as the smallest unit in the thermodynamic modelling process. As such, the complete Node Unit is the smallest feature that will be present in a simulation. The thermodynamic processes within a Node Group are only coarsely modelled as they are merely interesting for providing boundary conditions for the ComputeBox1 and ComputeBox2 simulations. The ComputeBox1 simulations will require – besides the arrangement of the Node Groups – the velocity field and temperature at the Node Group outlets over time as inbound boundary condition and will provide the room temperature over time at the outlet of the Node Group as outgoing boundary condition. Additionally, the heat generation of the PSU, switches, and other components have to be specified for a complete model of the ComputeBox1. For ComputeBox2, the same boundary conditions have to be defined as for the ComputeBox1 along with all other heat supplicants present in the room. Using the DEBB concept, the thermodynamic modelling can be tackled in a hierarchical way, reducing the complexity of the resulting overall model.

3.5 DEBB for configuration and reconfiguration

As mentioned in Section 3.1, a DEBB is described by characteristic data of its components and its sub-building blocks. The selection and configuration of components within a DEBB building block determines its energy efficiency and has to satisfy application and performance requirements.

Starting on Node-Level, a Node can be equipped with various **type** of main boards, CPUs, memory, network-interconnect, etc., each with specific energy consumption and performance behavior. A Node-Group, such as a RECS, can be equipped with various **types** and **number** of Nodes, each with different energy profile. In addition, the **organization** of different types of nodes within a Node-Group can be done in a certain way, i.e., utilizing/minimizing different thermal effects, such as a chimney effect. A ComputeBox1 (rack) can be equipped with various type and number of Nodes/Node-Groups, influencing its energy efficiency and density. The configuration and organization of Node-

Groups within the ComputeBox1 and selection of right integrated cooling approaches affects significantly the energy efficiency of a ComputeBox1 DEBB. A ComputeBox2 can be equipped with different type, density and number of ComputeBox1. The organization of several ComputeBox1 within a ComputeBox2 can be done in a certain way, i.e., finding the best air management strategy, as mentioned in 2.3.1. In addition, energy- and thermal-aware resource- and workload-management strategies, allowing reconfiguring resource settings at run-time (as mentioned in Section 2.5) have to be taken into consideration, to achieve higher energy efficiency. SVD-Toolkit, to be developed by WP2, will provide a toolset enabling to evaluate various configuration, reconfiguration and management strategies in simulations. This all has to be evaluated in various validation scenarios, as defined by WP6.

3.6 Assessing energy efficiency of DEBBs

As mentioned before, a (DEBB) concept describes energy efficiency of data-center building block on different levels of system granularity. A possible characterization of DEBBs in the terms of energy efficiency can be done according to Green Performance Indicators (GPIs), as defined in GAMES project, Jiang (2010) and Kipp et al. (2012). As described by Jiang (2010) and Kipp et al. (2012), GPIs are a measurement of the index of greenness of an IT system indicating the energy consumption, energy efficiency, energy saving potential and all energy related factors on different systems levels within IT service centre, including application and execution environment. In order to assess the global greenness of an application and IT infrastructure, GPIs were classified into four clusters: IT Resource Usage GPI, Application Lifecycle KPIs, Energy Impact GPIs and Organizational GPIs. Such a classification enables assessment of the energy efficiency of an IT centre from the business (organisational) level down to the technical level. GPIs allow also for considering, among others, the trade-off between performance and energy consumption at facility, application and compute node (IT infrastructure) level. As a DEBB is an abstraction of data centre building blocks on various levels of granularity, GPIs can be used for characterization of DEBBs on node level, node-group level, rack level and container level or entire IT centre. However, there is a need for extension of the GPIs to reflect influence of thermal conditions, allowing assessment of various cooling strategies in different environments.

The evaluation and extension of GPIs or other metrics for description and assessment of DEBBs is matter of ongoing research in CoolEmAll, in particular in WP5 and D5.1. The results of this research will be integrated in later Deliverables.

4 Conclusions

The above described concepts and techniques are a nearly full overview of what

is possible and state of the art in today's data centres as related to racks, servers, monitoring, controlling and cooling. From these concepts and techniques we have already chosen some to be integrated in the rack-level compute box (ComputeBox1) as explained in this Deliverable. Some details, for example the concrete cooling technique to integrate in the ComputeBox1, have still to be evaluated in the next months. For a concrete overview about the planned monitoring and controlling features see also the CoolEmAll Deliverable D4.1 where the architecture of the module operation platform (MOP) is introduced in which the monitoring and controlling approach is presented.

The DEBB concept introduced in this document will be refined in further Deliverables (D3.2), providing formal specification of formats as required and used by SVD Toolkit (WP2) to simulate thermal behavior of a DEBB. The selection of metrics used for the assessment of energy efficiency of DEBBs is matter of ongoing research in WP5, and will be integrated in future deliverables as well.

5 References

APC (2003). Air Distribution Architecture Options for Mission Critical Facilities.

APC (2005). Improving Rack Cooling Performance Using Blanking Panels. White Paper 44.

Barcelona Supercomputing Center [BSC] (2012). *MareNostrum System Architecture*. Refrieved from http://www.bsc.es/marenostrum-support-services/marenostrum-system-architecture

Christmann (2009). *Description for Resource Efficient Computing System (RECS)*. Retrieved from http://shared.christmann.info/download/project-recs.pdf

Cioara, T., Salomie, I., Anghel, I., Chira, J., Cocian, A., Henis, E. and Kat, R. (2011). *A Dynamic Power Management Controller for Optimizing Servers' Energy Consumption in Service Centers*. Lecture Notes in Computer Science, 2011, Volume 6568/2011, 158-168, DOI: 10.1007/978-3-642-19394-1_17

De Coufle, B. (2009). Effectiveness of Data Center Hot and Cold Aisle Containment.

DELL.COM/PowerSolutions (2009). *Precision Cooling for High Density Data CenterEnvironments.*

HP (2012). *HP Shapes the Future of Extreme Low-energy Server Technology*. Retrieved from http://www.hp.com/hpinfo/newsroom/press/2011/111101xa.html

IBM (2011). *Introduction to Blue Gene/Q*. Retrieved from http://public.dhe.ibm.com/common/ssi/ecm/en/dcl12345usen/DCL12345USEN.PDF

Jiang, T., Kipp, A., et al., Layered Green Performance Indicators Definitions, D2.1, Deliverable of GAMES project, 2010

Kennedy, D. *Inner-Rack Airflow Patterns & Data Center Efficiency.* Rittal White Paper 509

Khattar, M. K. (2010). Data Center Retrofit - Heat containment and Airflow Management. *ASHRAE Journal*

Kipp, A., Jiang, T., Fugini, M. et al. (2012) Layered Green Performance Indicators, 478-489. In *Future Generation Computer Systems* 28 (2).

Koomey, J. (2008). *Worldwide electricity used in data centers*. Environmental Research Letters. vol. 3, no. 034008.

Koomey, J., Berard, S., Sanchez, M, Wong, H (2010). Stanford University *Implications of Historical Trends in the Electrical Efficiency of Computing*. Annals of the History of Computing, IEEE, March 2011. Volume: 33 Issue 3, pp. 46-54.

Moore, G. E. (1965). *Cramming more components onto integrated circuits*. Electronics Magazine. p. 4.

Rasmussen, N. (2011). *Eco-mode: Benefits and Risks of Energy-saving Modes of UPS Operation*. Retrieved from http://www.apc.com/whitepaper?wp=157

Resch, M. (2011). *Growing Science at HLRS – Beyond Bare Metal*. inSiDE, Vol. 9 No. 2, Autumn 2011.

RITTAL. *Cold Aisle Containment for Improved Data Center Cooling Efficiency.* White Paper 506

Seamicro (2012). *SM10000-XE Highest Density, Most Energy-Efficient Xeon Server Ever Built*. Retrieved from http://www.seamicro.com/sm10000xe

Tozer, R. (2009). *Gestión del Aire en Centros de Cómputos - Principios.*

Tozer, R., Kurkjian, C., & Salim, M. (2009). *Air Management Metrics in Data Centers*. R. a.-C. American Society of Heating, ASHRAE Transactions, Vol. 115, part 1

Tyan (2012). *Supermicro 2U Twin^3 Solutions*. Retrieved from http://www.supermicro.com/products/nfo/2UTwin3.cfm

Volk, E. et al (2011). Towards Intelligent Management of Very Large Computing Systems, in proceedings of the CiHPC conference, 2010

US Department of Energy. *Data Center Rack Cooling with Rear-door Heat Exchanger.* Technology Case-Study Bulletin

A.S.H.R.A.E. (December 2011), *ASHRAE Journal, Data Center Environments, ASHRAE's Evolving Thermal Guidelines.* Technical Feature