Chapter 15
Energy-Aware Switch Design

Yukihiro Nakagawa  
Fujitsu Laboratories Ltd., Japan  

Osamu Shiraki  
Fujitsu Laboratories Ltd., Japan  

Takeshi Shimizu  
Fujitsu Laboratories Ltd., Japan  

Takashi Miyoshi  
Fujitsu Laboratories Ltd., Japan  

Takeshi Horie  
Fujitsu Laboratories Ltd., Japan  

Yasushi Umezawa  
Fujitsu Laboratories Ltd., Japan  

Yoichi Koyanagi  
Fujitsu Laboratories Ltd., Japan  

Akira Hattori  
Fujitsu Laboratories Ltd., Japan  

Yasuo Hidaka  
Fujitsu Laboratories of America, Inc., USA

ABSTRACT

The use of virtualization technology has been increasing in the IT industry to consolidate servers and reduce power consumption significantly. Virtualized commodity servers are scaled out in the data center and increase the demand for bandwidth between servers. Therefore, a high performance switch is required. The shared-memory switch is the best performance/cost switch architecture, but it is challenging to satisfy the requirements on the memory bandwidth in a high speed network. In addition, it is challenging to handle variable-length frames in Ethernet. This chapter describes the main challenges in Ethernet switch designs and then energy-aware switch designs, including switch architecture and high speed IO interface. As implementation examples, this chapter also describes a single-chip switch Large Scale Integration (LSI) embedded with high-speed IO interfaces and 10-Gigabit Ethernet (10GbE) switch blade equipped with the switch LSI. The switch blade delivers 100% more performance per watt than other 10GbE switch blades in the industry.

DOI: 10.4018/978-1-4666-1842-8.ch015
INTRODUCTION

Servers are great power consumers in the data center and often most of the servers are lightly loaded. This implies the major way to save power is to consolidate lightly loaded servers and replace them with virtual machines (Carr, 2008; Dawson & Bittman, 2008).

As a virtualization platform, a commodity server is used and it scales horizontally by increasing the number of servers in a data center (Cayton, 2008). The scale-out model increases the demand for bandwidth between servers inside a data center (Greenberg, Hamilton, & Jain, 2009). Therefore, a high performance network and switches are required. 10GbE is a practical interconnection technology as a high-speed, large-capacity network between servers, and a server and storage (IEEE 802.3 Working Group, 2010b). In addition, 10GbE technology enables Fibre Channel over Ethernet (FCoE) (ANSI, 2009) by removing the speed limit of 1GbE. FCoE consolidates Local Area Network (LAN) traffic and Storage Area Network (SAN) traffic, reducing the need for power and cooling.

This chapter describes the energy-aware switch designed by Fujitsu Laboratories for the high performance network. First, we describe the backgrounds and the main challenges of the switch design. Then we explain energy-aware switch designs including the switching architecture and high-speed IO interface. We also describe switch design considerations for emerging technologies such as FCoE and Edge Virtual Bridging (IEEE 802.1 Working Group, 2011c). We include implementations of a 10GbE switch LSI and a 10GbE switch blade for a large-scale blade server as an example of energy-aware switch designs.

BACKGROUND

The ideal network is one that directly connects computing resources with unlimited bandwidth and no latency. To approach this ideal, high-speed serial interconnects have become more common in communications environments, and, as a result, the role that switches play in these environments has become more important.

The shared-memory switch is the best performance/cost switch architecture (Chao & Liu, 2007). Memory has the best buffer utilization because it is shared by all inputs and outputs. In addition, delay performance is also the best because of no head-of-line blocking. On the other hand, the shared memory must have sufficient bandwidth to accept packets from all input ports and write packets to all output ports at the same time. In other words, the shared memory for a switch with N ports at a line rate R must have a bandwidth of 2*N*R at least. For example, the shared memory for a switch with 20 ports at a line rate 10Gb/s must have a bandwidth of 400Gb/s. If the memory access cycle is 10ns, the required data path is 4000bit wide which is not realistic. Several architectures have been proposed to satisfy the requirements on the memory bandwidth by using multiple shared-memory modules in parallel.

In the Space-Time-Space (STS)-type shared-memory switch, separate memories are shared among all input and output ports via switches (Oshima, Yamanaka, Saito, Yamada, Kohama, Kondoh, & Matsuda, 1992). Cells are written to the least occupied shared-memory first and the most occupied shared-memory last. This requires searching for the least occupied shared-memory. It may occur that two or more cells are read from the same shared-memory for different output ports. Therefore, this requires some kind of internal speedup to increase switch’s throughput.

In the Parallel Shared Memory (PSM) switch, separate memories are also shared among all input and output ports via switches (Iyer, Zhang, & McKeown, 2002). Cells are written to one of the shared-memories by analyzing “pigeon holes.” Each time slot, up to N packets or pigeons arrive which must immediately be placed in to a shared-memory or pigeon hole. Likewise, each