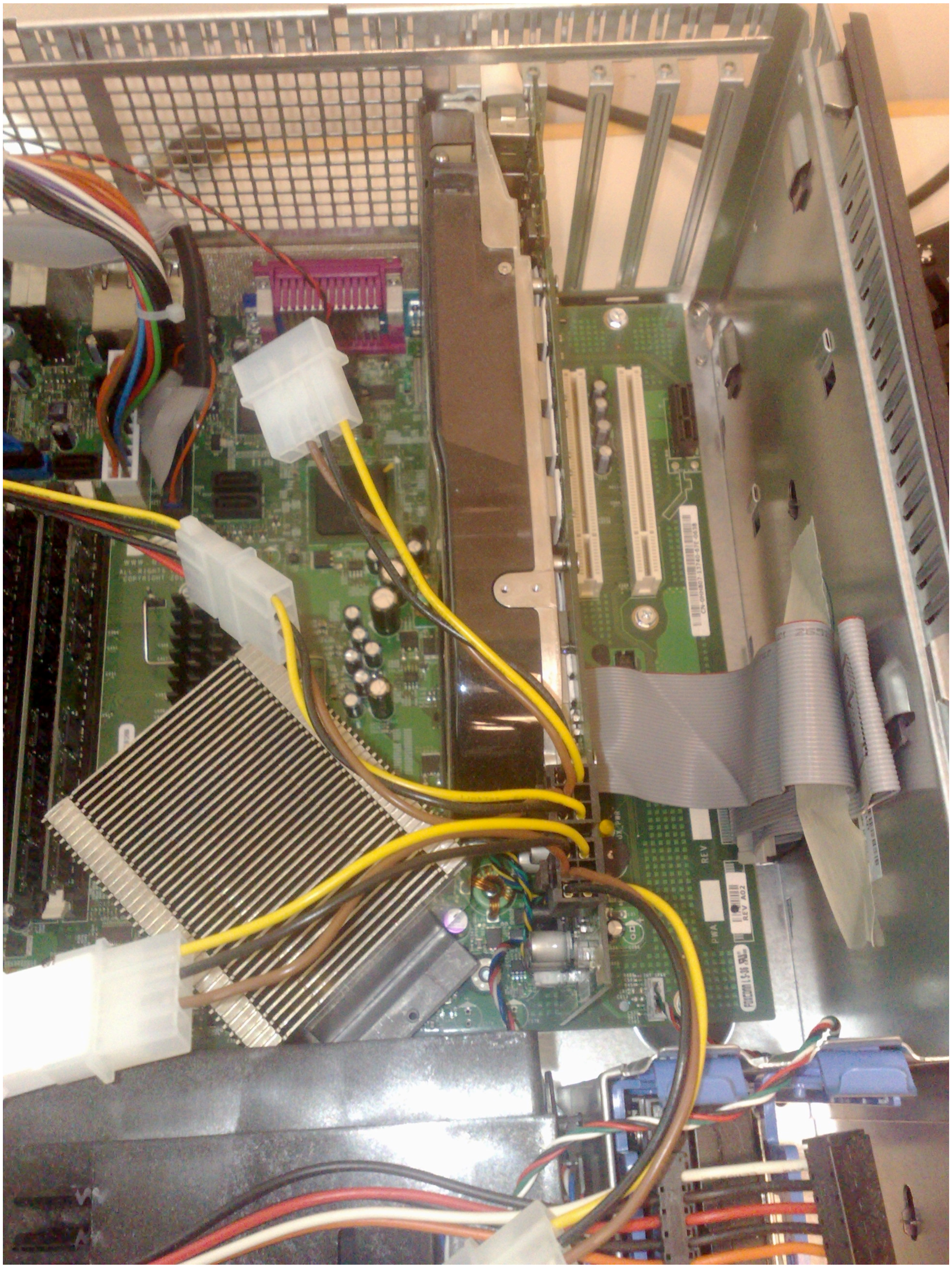

Figure 1: Building a PC for GPU computing—the GTX295 requires two slots and that the CPU cooler be twisted, but does work with the systems power supply.

Selecting the Right Hardware

Many *;login:* readers, me included, would not consider building a PC or installing a video card a particular challenge, but when it came to selecting the "right" components for GPU computing I had quite a different experience.

In the past, even more than for games, selecting the right video card would have a significant impact on the programming experience, while hardware compatibility was not a major issue. For example, the first search prototype based on graphics APIs required a fully programmable fragment shader, present only in cards supporting at least OpenGL 2.0. While theoretically openGL is supposed to emulate any features not present in the hardware, I was not able to run this prototype on any card older than a GeForce 6 series or on any other brand than NVIDIA.

Today low-end and even laptop video cards enable GPU computing, but only high-end video cards offer the latest features in terms of programmability. Besides the high price tag for faster and "better-looking" monsters, high-end cards come with strings attached, in particular increasing power consumption and their physical dimensions. While the first CUDA-capable card, the 8800GTX, already consumed 185W, two generations later, the top-of-the- line model, GTX295, requires close to 300W, more than most power supplies of standard office PCs provide for the whole system. Both cards are already one inch longer than a full-sized ATX mainboard, occupy two slots, and will not fit standard PC cases.

While I opted for a dedicated GPU compute server, whose specs looked like those of a gaming rig—2.6GHz Intel quad-core, 2GB of 1GHz RAM, and 2x 8800GTX—one of my colleagues decided to "simply" upgrade his machine with an 8800GTX video card and came up with a very unconventional solution. In order to fit the card in his desktop, he unscrewed the CPU cooler and rotated it by a few degrees. Since this made it impossible to affix the CPU to the mainboard, he placed his desktop sideways on top of his desk. He removed the side panel to increase air circulation, as the CPU was no longer fully covered by the cooler (Figure 1). After all this, his machine has been running stable for more than two years now, while over the same period I have replaced the mainboard, the memory, and one of the video cards of my GPU compute sever.

Although it is possible to run compute code on the same video card used for display, there are obvious performance implications and a time limit of five seconds. Thus our GPU compute server runs the console off a legacy PCI video card. However, this solution requires a mainboard that allows configuring the PCI card as the primary display and a PCI card supported by NVIDIA's CUDA-capable graphics driver, which would not install otherwise.

High-end mainboards have up to four PCI-e x16 slots, but this does not mean you can run four GPUs at x16 speeds. As opposed to the PCI bus, PCI-e implements multiple point–to-point connections, such that the total link speed is determined by the number of PCI-e lanes (*e.g.,* 36 for Intel's x58 chipset, allowing four x8 connections at most). While for small amounts of data (*e.g.,* a few thousand queries) this is not an issue, the time for transferring larger data sets increases linearly with size (Table 1).

data set size [MB]	1	2	4	8	16	32	64	128	256	512
transfer time [ms]	0.3	0.5	0.9	1.6	3.0	5.9	11.7	23.1	46.1	92.2

Table 1: Time required to copy data from main into video card memory on a x58 chipset and a GTX285.