Memory is a large component of computer system cost and current trends indicate this cost is increasing as a fraction of the total. Emerging applications such as in-memory databases, virtual machines, and big-data key-value stores demand more memory relative to compute. Some of these high-memory applications incidentally store many duplicate values in memory: in some cases, duplicates account for over 75% of the total. This talk describes a practical deduplicated memory that is compatible with existing hardware and software. Memory content from actual workloads indicates that deduplicated memory can effectively double your memory capacity. After describing a baseline memory design and reporting on its performance and bandwidth requirements, we next describe optimizations of this system that reduce memory traffic and improve performance relative to both the baseline deduplicated memory and, in many cases, relative to the original machine.

This was a derivative of the original HICAMP idea, which was acquired by Intel last year.

Bio

Mark Horowitz received his BS and MS in Electrical Engineering from MIT in 1978, and his PhD from Stanford in 1984. Since 1984 he has been a professor at Stanford working in the area of digital integrated circuit design. While at Stanford he has led a number of processor designs including: MIPS-X, one of the first processors to include an on-chip instruction cache; Torch, a statically-scheduled, superscalar processor; Flash, a flexible DSM machine; and Smash, a reconfigurable polymorphic manycore processor. He has also worked in a number of other chip design areas including high-speed memory design, high-bandwidth interfaces, and fast floating point. In 1990 he took leave from Stanford to help start Rambus Inc, a company designing high-bandwidth memory interface technology.