Systolic architectures, which permit multiple computations for each memory access, can speed execution of compute-bound problems without increasing I/O requirements.

Why Systolic Architectures?

H. T. Kung
Carnegie-Mellon University

High-performance, special-purpose computer systems are typically used to meet specific application requirements or to off-load computations that are especially taxing to general-purpose computers. As hardware cost and size continue to drop and processing requirements become well-understood in areas such as signal and image processing, more special-purpose systems are being constructed. However, since most of these systems are built on an ad hoc basis for specific tasks, methodological work in this area is rare. Because the knowledge gained from individual experiences is neither accumulated nor properly organized, the same errors are repeated. I/O and computation imbalance is a notable example—often, the fact that I/O interfaces cannot keep up with device speed is discovered only after constructing a high-speed, special-purpose device.

We intend to help correct this ad hoc approach by providing a general guideline—specifically, the concept of systolic architecture, a general methodology for mapping high-level computations into hardware structures. In a systolic system, data flows from the computer memory in a rhythmic fashion, passing through many processing elements before it returns to memory, much as blood circulates to and from the heart. The system works like an automobile assembly line where different people work on the same car at different times and many cars are assembled simultaneously. An assembly line is always linear, however, and systolic systems are sometimes two-dimensional. They can be rectangular, triangular, or hexagonal to make use of higher degrees of parallelism. Moreover, to implement a variety of computations, data flow in a systolic system may be at multiple speeds in multiple directions—both inputs and (partial) results flow, whereas only results flow in classical pipelined systems. Generally speaking, a systolic system is easy to implement because of its regularity and easy to reconfigure (to meet various outside constraints) because of its modularity.

The systolic architectural concept was developed at Carnegie-Mellon University,1,7 and versions of systolic processors are being designed and built by several industrial and governmental organizations.8-10 This article reviews the basic principle of systolic architectures and explains why they should result in cost-effective, high-performance special-purpose systems for a wide range of problems.

Key architectural issues in designing special-purpose systems

Roughly, the cycle for developing a special-purpose system can be divided into three phases—task definition, design, and implementation. During task definition, some system performance bottleneck is identified, and a decision on whether or not to resolve it with special-purpose hardware is made. The evaluation required for task definition is most fundamental, but since it is often application-dependent, we will concentrate only on architectural issues related to the design phase and will assume routine implementation.

Simple and regular design. Cost-effectiveness has always been a chief concern in designing special-purpose systems; their cost must be low enough to justify their limited applicability. Costs can be classified as recurring (design) and recurring (parts) costs. Part costs are dropping rapidly due to advances in integrated circuit technology, but this advantage applies equally to both special-purpose and general-purpose systems. Furthermore, since special-purpose systems are seldom produced in large quantities, part costs are less important than design costs. Hence, the design cost of a special-purpose system must be relatively small for it to be more attractive than a general-purpose approach.

Fortunately, special-purpose design costs can be reduced by the use of appropriate architectures. If a structure can truly be decomposed into a few simple substructures of building blocks, which are used repetitively with simple interfaces, great savings can be achieved. This is especially true for VLSI design where a single chip comprises hundreds of thousands of components. To cope with this complexity, simple and regular designs, similar
Concurrency and communication. There are essentially two ways to build a fast computer system. One is to use fast components, and the other is to use concurrency. The last decade has seen an order of magnitude decrease in the cost and size of computer components but only an incremental increase in component speed. With current technology, tens of thousands of gates can be put in a single chip, but no gate is much faster than its TTL counterpart of 10 years ago. Since the technological trend clearly indicates a diminishing growth rate for component speed, any major improvement in computation speed must come from the concurrent use of many processing elements. The degree of concurrency in a special-purpose system is largely determined by the underlying algorithm. Massive parallelism can be achieved if the algorithm is designed to introduce high degrees of pipelining and multiprocessor. When a large number of processing elements work simultaneously, coordination and communication become significant—especially with VLSI technology where routing costs dominate the power, time, and area required to implement a computation. The issue here is to design algorithms that support high degrees of concurrency, and in the meantime to employ simple, regular communication and control to enable efficient implementation.

Balancing computation with I/O. Since a special-purpose system typically receives data and outputs results through an attached host, I/O considerations influence overall performance. (The host in this context can mean a computer, a memory, a real-time device, etc. In practice, the special-purpose system may actually input from one "physical" host and output to another.) The ultimate performance goal of a special-purpose system is—and should be no more than—a computation rate that balances the available I/O bandwidth with the host. Since an accurate a priori estimate of available I/O bandwidth in a complex system is usually impossible, the design of a special-purpose system should be modular so that its structure can be easily adjusted to match a variety of I/O bandwidths.

Suppose that the I/O bandwidth between the host and the special-purpose system is 10 million bytes per second, a rather high bandwidth for present technology. Assuming that at least two bytes are read or written to the special-purpose system for each operation, the maximum rate will be only 5 million operations per second, no matter how fast the special-purpose system can operate (see Figure 1). Orders of magnitude improvements in this throughput are possible only if multiple computations are performed per I/O access. However, the repetitive use of a data item requires that it be stored inside the system for a sufficient length of time. Thus, the I/O problem is related not only to the available I/O bandwidth, but also to the available memory latencies to the system. The question then is how to arrange a computation together with an appropriate memory structure so that computation time is balanced with I/O time.

The I/O problem becomes especially severe when a large computation is performed on a small special-purpose system. In this case, the computation must be decomposed into subcomputations. Executing one subcomputation at a time may require a substantial amount of I/O to store or retrieve intermediate results. Consider, for example, performing the n-point fast Fourier transform using an S-point device when n is large and S is small. Figure 2 depicts the n-point FFT computation and a decomposition scheme for n = 16 and S = 4. Note that each subcomputation block is sufficiently small so that it can be handled by the 4-point device. During execution, the results of a block must be temporarily stored to the host and later retrieved to be combined with the results of other blocks as they become available. With the decomposition scheme shown in Figure 2b, the total number of I/O operations is O(n log n / log S). In fact, it has been shown that to perform the n-point FFT with a device of O(S) memory, at least as many I/O operations are needed for any decomposition scheme. Thus, for the n-point FFT problem, a 5-point device cannot achieve more than an O(log S) speed-up ratio over the conventional O(n log n) software implementation time, and since it is a consequence of the I/O concern, this upper bound holds independent of device speed. Similar upper bounds have been established for speed-up ratios achievable by devices for other computations such as sorting and matrix multiplication.

Knowing the I/O-imposed performance limit helps prevent overkill in the design of a special-purpose device. In practice, problems are typically "larger" than special-purpose devices. Therefore, questions such as how a computation can be decomposed to minimize I/O, how the I/O requirement is related to the size of a special-purpose system and its memory, and how the I/O bandwidth limits the speed-up ratio achievable by a special-purpose system present another set of challenges to the system architect.
Systolic architectures: the basic principle

As a solution to the above challenges, we introduce systolic architectures, an architectural concept originally proposed for VLSI implementation of some matrix operations. Examples of systolic architectures follow in the next section, which contains a walk-through of a family of designs for the convolution computation.

A systolic system consists of a set of interconnected cells, each capable of performing some simple operation. Because simple, regular communication and control structures have substantial advantages over complicated ones in design and implementation, cells in a systolic system are typically interconnected to form a systolic array or a systolic tree. Information in a systolic system flows between cells in a pipelined fashion, and communication with the outside world occurs only at the “boundary cells.” For example, in a systolic array, only those cells on the array boundaries may be I/O ports for the system.

Computational tasks can be conceptually classified into two families—compute-bound computations and I/O-bound computations. In a computation, if the total number of operations is larger than the total number of input and output elements, then the computation is compute-bound, otherwise it is I/O-bound. For example, the ordinary matrix-matrix multiplication algorithm represents a compute-bound task, since every entry in a matrix is multiplied by all entries in some row or column of the other matrix. Adding two matrices, on the other hand, is I/O-bound, since the total number of adds is not larger than the total number of entries in the two matrices.

It should be clear that any attempt to speed up an I/O-bound computation must rely on an increase in memory bandwidth. Memory bandwidth can be increased by the use of either fast components (which could be expensive) or interleaved memories (which could create complicated memory management problems). Speeding up a compute-bound computation, however, may often be accomplished in a relatively simple and inexpensive manner, that is, by the systolic approach.

The basic principle of a systolic architecture, a systolic array in particular, is illustrated in Figure 1. By replacing a single processing element with an array of PEs, or cells in the terminology of this article, a higher computation throughput can be achieved without increasing memory bandwidth. The function of the memory in the diagram is analogous to that of the heart; it “pumps” data (instead of blood) through the array of cells. The crux of this approach is to ensure that once a data item is brought out from the memory it can be used effectively at each cell it passes while being “pumped” from cell to cell along the array. This is possible for a wide class of compute-bound computations where multiple operations are performed on each data item in a repetitive manner.

Being able to use each input data item a number of times (and thus achieving high computation throughput with only modest memory bandwidth) is just one of the many advantages of the systolic approach. Other advantages, such as modular expandability, simple and regular data and control flows, use of simple and uniform cells, elimination of global broadcasting, and fan-in (and possibly) fast response time, will be illustrated in various systolic designs in the next section.

A family of systolic designs for the convolution computation

To provide concrete examples of various systolic structures, this section presents a family of systolic designs for the convolution problem, which is defined as follows:

Given the sequence of weights \( w_1, w_2, \ldots, w_k \)
and the input sequence \( x_1, x_2, \ldots, x_n \)

compute the result sequence \( y_1, y_2, \ldots, y_{n-k} \) defined by

\[
y_i = w_i x_i + w_{i+1} x_{i+1} + \cdots + w_k x_{i+k-1}
\]

Figure 2. (a) 16-point fast-Fourier-transform graph; (b) decomposing the FFT computation with \( n = 16 \) and \( S = 4 \).
We consider the convolution problem because it is a simple problem with a variety of enlightening systolic solutions, because it is an important problem in its own right, and more importantly, because it is representative of a wide class of computations suited to systolic designs. The convolution problem can be viewed as a problem of combining two data streams, \( w_i \)'s and \( x_i \)'s, in a certain manner (for example, as in the above equation) to form a resultant data stream of \( y_i \)'s. This type of computation is common to a number of computation routines, such as filtering, pattern matching, correlation, interpolation, polynomial evaluation (including discrete Fourier transforms), and polynomial multiplication and division. For example, if multiplication and addition are interpreted as comparison and boolean AND, respectively, then the convolution problem becomes the pattern matching problem.1 Architectural concepts for the convolution problem can thus be applied to other problems as well.

The convolution problem is compute-bound, since each input \( x_i \) is to be multiplied by each of the \( k \) weights. If the \( x_i \) is input separately from memory for each multiplication, then when \( k \) is large, memory bandwidth becomes a bottleneck, precluding a high-performance solution. As indicated earlier, a systolic architecture resolves this I/O bottleneck by making multiple use of each \( x_i \) fetched from the memory. Based on this principle, several systolic designs for solving the convolution problem are described below. For simplicity, all illustrations assume that \( k = 3 \).

(Semi-)systolic convolution arrays with global data communication. If an \( x_i \), once brought out from the memory, is broadcast to a number of cells, then the same \( x_i \) can be used by all the cells. This broadcasting technique is probably one of the most obvious ways to make multiple use of each input element. The opposite of broadcasting is fan-in, through which data items from a number of cells can be collected. The fan-in technique can also be used in a straightforward manner to resolve the I/O bottleneck problem. In the following, we describe systolic designs that utilize broadcasting and fan-in.

**Design B1**—broadcast inputs, move results, weights stay. The systolic array and its cell definition are depicted in Figure 3. Weights are preloaded to the cells, one at each cell, and stay at the cells throughout the computation. Partial results \( y_i \) move systolically from cell to cell in the left-to-right direction, that is, each of them moves over the cell to its right during each cycle. At the beginning of a cycle, one \( x_i \) is broadcast to all the cells and one \( y_i \), initialized as zero, enters the left-most cell. During cycle \( i \), \( w_i \cdot x_i \) is accumulated to \( y_i \), and the resulting sum \( y_i \) is moved to the cell to the right. During cycle \( i+1 \), \( w_i \cdot x_i \) is accumulated to \( y_i \), and then \( y_i \) is moved to the cell to the right. This process continues until all the partial results have moved through the array and the final results are collected in the cell to the right of the last cell. The basic principle of this design was previously proposed for circuits to implement a pattern matching processor10 and for circuits to implement polynomial multiplication.17 20

**Design B2**—broadcast inputs, move weights, results stay. In design B2 (see Figure 4), each \( y_i \) stays at a cell to accumulate its terms, allowing efficient use of available multiplier-accumulator hardware. Indeed, this design was described in an application booklet for the TRW multiplier-accumulator chips.21 The weights circulate around the array of cells, and the first weight \( w_i \) is associated with a tag bit that signals the accumulator to output and reset its contents. In Design B1 (Figure 3), the systolic path for moving \( y_i \)'s may be considerably wider than that for moving \( w_i \)'s in design B2 because for numerical accuracy \( y_i \)'s typically carry more bits than \( w_i \)'s. The use of multiplier-accumulators in design B2 may also help increase precision of the results, since extra bits can be kept in these accumulators with modest cost. Design B1, however, does have the advantage of not requiring a separate bus (or other global network), denoted by a dashed line in Figure 4, for collecting outputs from individual cells.

**Design F**—fan-in results, move inputs, weights stay. If we consider the vector of weights \( (w_1, w_2, \ldots, w_k) \) being fixed in space and input vector \( (x_1, x_2, \ldots, x_k) \) as sliding over the weights in the left-to-right direction, then the convolution problem is one that computes the inner product of the weight vector and the section of input vector it overlaps. This view suggests the systolic array is discussed later.

**Design R1**—resulting product, move inputs, opposite directions. In this initial result, \( y_i \) stays at a cell and \( w_i \)'s move systolically when an \( x_i \) meets a \( w_i \). The resulting product is a cell. To ensure that each successive \( x_i \)'s on the \( x \) axis overlap with the \( y_i \)'s already there, \( x_i \)'s on the same cycle times and so are the \( y_i \)'s.

Like design B2, design R1 takes advantage of multiplier-accumulator chips by using a tag bit associated with each cell to determine the output and reset the cell.
Figure 5. Design F: systolic convolution array (a) and cell (b) where $w_i$'s move systolically, and $y_i$'s are formed through the fan-in of results from all the cells.

Figure 6. Design R1: systolic convolution array (a) and cell (b) where $y_i$'s stay and $x_j$'s and $y_j$'s move in opposite directions systolically.

Figure 7. Design R2: systolic convolution array (a) and cell (b) where $y_i$'s stay and $x_j$'s and $w_i$'s both move in the same direction but at different speeds.
each $w_j$ stays inside every cell it passes for one extra cycle, thus taking twice as long to move through the array as any $x_i$. In this design, multiplier-accumulator hardware can be used effectively and so can the tag bit method to signal the output of the accumulator contents at each cell. Compared to design R1, this design has the advantage that all cells work all the time when performing a single convolution, but it requires an additional register in each cell to hold a $w$ value. This algorithm has been used for implementing a pipeline multiplier.

There is a dual version of design R2; we can have the $w_j$ move twice as fast as the $x_i$. To create delays for the $x$ data stream, this dual design requires a register in each cell for storing an $x$ rather than a $w$ value. For circumstances where the $w$'s carry more bits than the $x$'s, the dual design becomes attractive.

**Design W1**—weights stay, inputs and results move in opposite directions. In design W1 (and design W2, below), weights stay, one at each cell, but results and inputs move systolically. These designs are not geared to the most effective use of available multiplier-accumulator hardware, but for some other circumstances they are potentially more efficient than the other designs. Because the same set of weights is used for computing all the $y_j$'s and different sets of the $x_i$ are used for computing different $y_j$'s, it is natural to have the $w_j$ preloaded to the cells and stay there, and let the $x_i$'s and the $y_j$'s move along the array. We will see some advantages of this arrangement in the systolic array depicted in Figure 8, which is a special case of a proposed systolic filtering array. This design is fundamental in the sense that it can be naturally extended to perform recursive filtering and polynomial division.

In design W1, the $w_j$'s stay and the $x_i$'s and $y_j$'s move systolically in opposite directions. Similar to design R1, consecutive $x_i$'s and $y_j$'s are separated by two cycle times. Note that because the systolic path for moving the $y_j$'s already exists, there is no need for another systolic output path as in designs R1 and R2. Furthermore, for each $i$, $y_j$ outputs from the left-most cell during the same cycle as its last input, $x_{i+k}$ (or $x_{i-k}$ for $k=1$), enters that cell. Thus, this systolic array is capable of outputting a $y_j$ every two cycle times with constant response time. Design W1, however, suffers from the same drawback as design R1, namely, only approximately one-half the cells work at any given time unless two independent convolution computations are interleaved in the same array. The next design, like design R2, overcomes this shortcoming by having both the $x_i$'s and $y_j$'s move in the same direction but at different speeds.

**Design W2**—weights stay, inputs and results move in the same direction but at different speeds. With design W2 (Figure 9), all the cells work all the time, but it loses the advantage of design W1, the constant response time. The output of $y_i$ now takes place $F$ cycles after the last of its inputs starts entering the left-most cell of the systolic array. This design has been extended to implement 2-D convolutions, where high throughputs rather than the response times are of concern. Similar to design R1, design W2 has a dual version for which the $x_i$'s move twice as fast as the $y_j$'s.

**Remarks.** The designs presented above have not exhausted all the possible systolic designs for the convolution problem. For example, it is possible to have systolic arrays where the weights, weights, and inputs all move during each cycle. It could also be advantageous to include into each cell a "cell memory" capable of storing a set of weights. With this feature, using a systolic control (or address) path, weights can be selected on the fly to implement convolution or adaptive filtering. Moreover, flexibility introduced by the cell memories and systolic control can make the same systolic array implement different functions. Indeed, the ESI systolic processor utilizes cell memories to implement multiple functions including convolution and matrix multiplication.

Once one systolic design is obtained for a problem, it is likely that a set of other systolic designs can be derived similarly. The challenge is to understand precisely how strengths and drawbacks of each design so that a more appropriate design can be selected for a given environment. For example, if there are more weights than cells, it is useful to know that a scheme where partial results are generally requires less I/O than one where partial results move, since the latter scheme requires partial results to input and output many times. A single-multiplier-accumulator hardware component often represents a critical implementation of the multiplier and adder needed by each cell. However, for implementations to be worthwhile, it is important to allow overlap of cells. To achieve this, the implementation of some semi-systolic designs R1 and B require more complex systolic arrays.

**Criteria and admissibility.**

Having described systolic arrays, we can now begin to consider criteria for their admissibility.

1. _The design must be admissible._ Because of the high throughput, for outside communication, the systolic array can either use global broadcast and unidirectional links or traverse through an array cell. For modular designs, the second approach is preferable.
2. _The design uses the power of a systolic design._ It is not sufficient to have many simple and powerful processors. Concurrent processing of parallel instructions is useful but not sufficient. The problem is to use a few powerful processors. Success depends on the stages involved in the result (for example, the design results in parallel) and on some design, such as the overlap I/O and concurrency and parallelism.
Figure 10. Overlapping the executions of multiply and add in design W1.

Figure 11. Two-dimensional systolic arrays: (a) type R, (b) type H, and (c) type T.

Criteria and advantages

Having described a family of systolic convolution arrays, we can now be more precise in suggesting and evaluating criteria for the design of systolic structures.

1. The design makes multiple use of each input data item. Because of this property, systolic systems can achieve high throughputs with modest I/O bandwidths for outside communication. To meet this criterion, one can either use global data communications, such as broadcast and unbounded fan-in, or have each input travel through an array of cells so that it is used at each cell. For modular expansibility of the resulting system, the second approach is preferable.

2. The design uses extensive concurrency. The processing power of a systolic architecture comes from concurrent use of many simple cells rather than sequential use of a few powerful processors as in many conventional architectures. Concurrency can be obtained by pipelining the stages involved in the computation of each single result (for example, design B1), by multiprocessing many results in parallel (designs R1 and R2), or by both. For some designs, such as W1, it is possible to completely overlap I/O and computation times to further increase concurrency and provide constant-time responses.

To a given problem there could be both one- and two-dimensional systolic array solutions. For example, two-dimensional convolution can be performed by a one-dimensional systolic array or a two-dimensional systolic array. When the memory speed is more than cell speed, two-dimensional systolic arrays such as those depicted in Figure 11 should be used. At each cell cycle, all the I/O ports on the array boundaries can input or output data items to or from the memory; as a result, the available memory bandwidth can be fully utilized. Thus, the choice of a one- or two-dimensional scheme is very dependent on how cells and memories will be implemented.

As in one-dimensional systolic array, data in two-dimensional arrays may flow in multiple directions and at multiple speeds. For examples of two-dimensional systolic array, see Guibas et al. and Kung and Leiserson (type R), Kung and Leiserson and Weiser and Davis (type H), and Bojanczyk et al. and Gentleman and Kung (type T). In practice, systolic arrays can be chained together to form powerful systems such as the one depicted in Figure 12, which is capable of producing on-the-fly least-squares fits to all the data that have arrived up to any given moment.

For the systolic structures discussed in the preceding section, computations are pipelined over an array of cells. To permit even higher concurrency, it is sometimes possible to introduce another level of pipelining by allowing the operations inside the cells themselves to be pipelined. (Note that pipelined arithmetic units will become increasingly common as VLSI makes the extra circuits needed for
staging affordable.) Both designs W1 and W2 support two-level pipelining. Since system cycle time is the time of a stage of a cell, rather than the whole cell cycle time, two-level pipelined systolic systems significantly improve throughput.

3) There are only a few types of simple cells. To achieve performance goals, a systolic system is likely to use a large number of cells. The cells must be simple and of only a few types to curtail design and implementation costs, but exactly how simple is a question that can only be answered on a case by case basis. For example, if a systolic system consisting of many cells is to be implemented on a single chip, each cell should probably contain only simple logic circuits plus a few words of memory. On the other hand, for board implementation each cell could reasonably contain a high-performance arithmetic unit plus a few thousand words of memory. There is, of course, always a trade-off between cell simplicity and flexibility.

4) Data and control flows are simple and regular. Pure systolic systems totally avoid long-distance or irregular wires for data communication. The only global communication (besides power and ground) is the system clock. Of course, self-timed schemes can be used instead for synchronizing neighboring cells, but efficient implementations of self-timed protocols may be difficult. Fortunately, for any one-dimensional systolic array, a global clock parallel to the array presents no problems, even if the array is arbitrarily long. The systolic array (with data flowing in either one or opposite directions) will operate correctly despite the possibility of a large clock skew between its two ends. However, large two-dimensional arrays may require slowdown of the global clock to compensate for clock skew. Except for this possible problem in the two-dimensional case, systolic designs are completely modular and expandable; they present no difficult synchronization or resource conflict problems. Software overhead associated with operations such as address indexing are totally eliminated in systolic systems. This advantage alone can mean a substantial performance improvement over conventional general-purpose computers. Simple, regular control and communication also imply simple, area-efficient layout or wiring—an important advantage in VLSI implementation.

In summary, systolic designs based on these criteria are simple (a consequence of properties 3 and 4), modular and expandable (property 4), and yield high performance (properties 1, 2, and 4). They therefore meet the architectural challenges for special-purpose systems. A unique characteristic of the systolic approach is that as the number of cells expands the system cost and performance increase proportionally, provided that the size of the underlying problem is sufficiently large. For example, a systolic convolution array can use an arbitrarily large number of cells cost-effectively, if the kernel size (that is, the number of weights) is large. This is in contrast to other parallel architectures which are seldom cost-effective for more than a small number of processors. From a user's point of view, a systolic system is easy to use—he simply several of these pumps in the input data and then receives the results either designating systolic on-the-fly or at the end of the computation.

Summary and concluding remarks

Bottlenecks to speeding up a computation are often to limited system memory bandwidths, called von Neumann bottlenecks, rather than limited processing capabilities per se. This problem can certainly be expected for I/O-bound computations, but with a conventional architectural approach, it may be present even for compute-bound computations. For every operation, at least one operand has to be fetched (or stored) from or to memory, so the total amount of I/O is proportional to the number of operations rather than the number of inputs and outputs. Thus, a problem that was originally compute-bound can become I/O-bound during its execution.

This unfortunate situation is the result of a mismatch between the computation and the architecture. Systolic architectures, which ensure multiple computations to memory access, can speed up compute-bound computations without increasing I/O requirements.

The convolution problem is just one of many computations that can benefit from the systolic approach. Systolic designs using one- or two-dimensional array or tree structures are available for following regular, compute-bound computations.

Signal and image processing:
- FIR, IIR filtering, and 1-D convolution
- 2-D convolution
- discrete Fourier transforms
- interpolation
- 1-D and 2-D geometric warping
problems, even if the array (with data flow-
ions) will operate cor-
gec observable human to compensate
problem in the desig
are completely no difficulty
Software
ations such as add
systolic systems. This ad-
tial performance in
general-purpose com-
unication also at or wiring—an impor-
tation.
ated on these criteria are,
modules, yield high perfor-
before meet the architec-
ose systems. A unique
such as the com-
est performance but that the size of the
ible for applications. For example, a use
arily large if the kernel size that is, is in
(formatter) to such
U.S. easy to use—he simply
ceives the results either

i remarks

computation are often due widths, so called von Neum-
limited processing cap-
ically be expected for
conventional archi-
for computation, at least one or
do (or stored) from (or to)
O is proportional to the
the number of inputs
was originally con-
run during its execution.
result of a mismatch be-
architecture. Systolic ar-
threshold computations per

computation 2,3,11;
• 2-D convolution and correlation 6,8,10,24,25;
• discrete Fourier transform 2,3;
• interpolation 24;
• 1-D and 2-D median filtering 32; and
• geometric warping 24.

Matrix arithmetic:
• matrix-vector multiplication 5;
• matrix-matrix multiplication 5,27;
• matrix triangularization (solution of linear systems, matrix inversion) 3,29;
• QR decomposition (eigenvalue, least-square computations) 28,29; and
• solution of triangular linear systems 5.

Non-numeric applications:
• data structures—stack and queue, 24 searching, 15,25,26 priority queue, 1 and sorting 1,15;
• graph algorithms—transitive closure, 26 minimum spanning trees, 37 and connected components 18;
• language recognition—string matching and regular expression 59;
• dynamic programming 26;
• encoders (polynomial division) 23; and
• relational data-base operations 4,40.

In general, systolic designs apply to any compute-
end problem that is regular—that is, one where repeti-
tions are performed on a large set of data. This, the above list is certainly not complete (and was not
intended to be so). Its purpose is to provide a range of
examples and possible applications. After studying
several of these examples, one should be able to start
designing systolic systems for one's own tasks. Some
systolic solutions can usually be found without too much
difficulty. (I know of only one compute-bound problem
that arises naturally in practice for which no systolic solu-
tion is known, and I cannot prove that a systolic solution
is impossible.) This probably due to the fact that most
compute-bound problems are inherently regular in the
sense that they are definable in terms of simple recur-
ences. Indeed, the notion of systolicity is implicit in quite
a few previously known special purpose designs, such as
the sorting 41 and multiply designs 22. This should not
come as a surprise; as we have been arguing systolic struc-
tures are essential for obtaining any cost-effective, high-
performance solution to compute-bound problems. It is
useful, however, to make the systolic concept explicit so
that designers will be conscious of this important design
criterion.

While numerous systolic designs are known today, the
question of their automatic design is still open. But recent
efforts show significant progress. 27,42 Leiserson and
Saxe, for instance, can convert some semi-systolic
systems involving broadcasting or unbounded fan-in into
pure-systolic systems without global data communication.
A related open problem concerns the specification and
verification of systolic structures. For implementa-
tion and proof purposes, rigorous notation other than in-
formal pictures (as used in this article) for specifying
systolic designs is desirable.

With the development of systolic architectures, more
and more special-purpose systems will become feasi-
ble—especially systems that implement fixed, well-under-
stood computation routines. But the ultimate goal is ef-
effective use of systolic processors in general computing en-
vironments to off-load regular, compute-bound compu-
tations. To achieve this goal further research is needed in
two areas. The first concerns the system integration: we
must provide a convenient means for incorporating high-
performance systolic processors into a complete system
and for understanding their effective utilization from a
system point of view. The second research area is to
specify building-blocks for a variety of systolic processors
so that, once built, these building blocks can be pro-
grammed to form basic cells for a number of systolic
systems. The building-block approach seems inherently
suitable to systolic architectures since they tend to use only
a few types of simple cells. By combining these build-

References


