

Evaluating polynomials in several variables and their derivatives on a GPU computing processor*

Jan Verschelde and Genady Yoffe
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
851 South Morgan (M/C 249)
Chicago, IL 60607-7045, USA
`jan@math.uic.edu`, `gyoffe2@uic.edu`
`www.math.uic.edu/~jan`, `www.math.uic.edu/~gyoffe`

28 February 2012

Abstract

In order to obtain more accurate solutions of polynomial systems with numerical continuation methods we use multiprecision arithmetic. Our goal is to offset the overhead of double double arithmetic accelerating the path trackers and in particular Newton's method with a general purpose graphics processing unit. In this paper we describe algorithms for the massively parallel evaluation and differentiation of sparse polynomials in several variables. We report on our implementation of the algorithmic differentiation of products of variables on the NVIDIA Tesla C2050 Computing Processor using the NVIDIA CUDA compiler tools.

Key words and phrases. algorithmic differentiation, compute unified device architecture (CUDA), graphics processing unit (GPU), massively parallel polynomial evaluation, Speelpenning product.

1 Introduction

The problem we consider in this paper is the efficient evaluation of a polynomial system and its Jacobian matrix on a graphics processing unit (GPU), the NVIDIA Tesla C2050 Computing Processor. For an introduction to GPU computing, we refer to [17] and [29]. The success of general purpose GPU computing in many areas of scientific computing is explained in [16].

The evaluation of a polynomial system and its Jacobian matrix is a computationally intensive stage in Newton's method to approximate an isolated solution. Numerical continuation methods apply Newton's method as corrector in predictor-corrector algorithms to track paths of solutions defined by a homotopy. Homotopy continuation methods have led to efficient numerical solvers of polynomial systems (see e.g. [2], [23], [28], [42]) and constitute the computational engine in the emerging area of numerical algebraic geometry ([19], [34], [35]).

*This material is based upon work supported by the National Science Foundation under Grant No. 1115777.

Granularity issues and parallel complexity of continuation methods for nonlinear systems are discussed in [3] and [6]. If one is interested in computing *all* isolated solutions of a polynomial system, then distributing path tracking jobs in a manager/worker paradigm using message passing [33] leads to very good speedups. Such parallel implementations are in **Bertini** [4], **HOM4PS-2.0para** [24], **PHoMpara** [14], **POLSYS_GLP** [36], and **PHCpack** [37], documented in [13], [20, 21], [22], [38], and [41].

For large polynomial systems in many variables and of high degrees we have observed that (1) the cost of polynomial evaluation often dominates the cost of linear algebra operations; and (2) the double precision in standard hardware is often insufficient to guarantee accurate results. When running many path tracking jobs, just one solution path may require extended multiprecision arithmetic. Then we need to apply parallel algorithms to offset the extra cost of software driven arithmetic. In analogy to speedup, we use the notion of *quality up* (inspired by [1]) and ask the question: given p processors (or cores) how much extra precision can we afford in roughly the same time as a sequential run?

Because of simple memory management on shared memory multicore processing we selected the quad double library **QD 2.3.9** [15]. The ideas to achieve extended precision using hardware doubles originate in [7], see also [30], [31] and [32]. In [40], we determined experimentally that the cost factor in the overhead of using double double arithmetic is around 8, coinciding with the number of cores on the workstation we were using at that time. Then the cost of tracking one solution path in double double arithmetic can be compensated in a parallel multicore implementation, thus achieving quality up. Using techniques of algorithmic differentiation [12], we extended this work in [39].

This current paper describes our efforts to offset the cost of extended precision by a parallel implementation of evaluation and differentiation algorithms using double double and quad double precision arithmetic on a GPU. We refer to [25] for a report on quad doubles on a GPU.

Related work in algebraic computations on a GPU are polynomial multiplication [8], [26], resultant [9], GCD computations [10], and solving bivariate polynomial systems [27]. Parallel automatic differentiation techniques are described in [5] and [11].

2 Problem Statement

The problem we consider is the evaluation of system of polynomial equations in several variables and all its derivatives (as needed in the Jacobian matrix of the system). Let n denote the number of variables. A polynomial f in n variables $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is stored as a tuple (C, A) of complex coefficients C and corresponding exponents A . In multi-index notation we write f as

$$f(\mathbf{x}) = \sum_{\mathbf{a} \in A} c_{\mathbf{a}} \mathbf{x}^{\mathbf{a}}, c \in \mathbb{C} \setminus \{0\}, \mathbf{x}^{\mathbf{a}} = x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}. \quad (1)$$

Then a system $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, $\mathbf{f} = (f_1, f_2, \dots, f_n)$, is defined by a tuple of coefficients and exponents.

In our problem setup, we consider as inputs *sparse* polynomials, that is: only relatively few monomials appear with nonzero coefficients, few relative to the degree of the polynomials. Typically we take the number of monomials as $O(n)$. For dense polynomials, a nested Horner scheme is recommended, see [18]. Because of the exponential growth of the number of monomials,

dense polynomials in several variables of high degree do not occur often in applications. For establishing benchmarks we consider in this paper systems with a fixed number k of variables in monomials, a fixed maximal degree d up to which any of variables can appear in monomials of the system, and a fixed number m of monomials in all polynomials.

There are three stages in the evaluation of a polynomial system and its Jacobian matrix:

1. The computation of the monomial products $x_{i_1}^{a_{i_1}-1} x_{i_2}^{a_{i_2}-1} \cdots x_{i_k}^{a_{i_k}-1}$ where the exponents $a_{i_j} \geq 1$, $j = 1, 2, \dots, k$, and $1 \leq i_1 < i_2 < \dots < i_k \leq n$. A preprocessing step is the computation of all powers of x_i^j , for all $i \in \{1, 2, \dots, n\}$ and $j \in \{2, \dots, d-1\}$.
2. The evaluation of products of variables $x_{i_1} x_{i_2} \cdots x_{i_k}$, and all their derivatives. The product of variables $x_{i_1} x_{i_2} \cdots x_{i_k}$ is known in [12, page 3] as the example of Speelpenning. We call this product of variables a *Speelpenning product*. In this stage we also multiply Speelpenning products and their derivatives with the values of monomial products computed in the previous stage to obtain the values of monomials and the monomial parts of their derivatives.
3. The multiplication of the coefficients with the corresponding evaluated monomials, followed by summations of obtained products to get the values of the polynomials of the system and the Jacobian.

While the evaluation of high dimensional and high degree polynomials is computationally intensive, the challenge for data parallelism lies in the irregularity of the data. In order to achieve good speedups, we will derive regularity assumptions on the input data.

3 Massively Parallel Algorithms

Following the three stages (outlined in the problem statement section) in the evaluation of a polynomial system and its Jacobian matrix, we devote one subsection to each stage. There are three kernels. The first kernel corresponds to the first stage, whereas the second kernel is described in subsections 2 and 3 below. We explain the third kernel in the second half of the third subsection.

3.1 Common Factor Calculation

To evaluate a monomial $x_1^3 x_2^7 x_3^2$ and its derivatives, we first evaluate the factor $x_1^2 x_2^6 x_3$ and then multiply this factor with all derivatives of $x_1 x_2 x_3$. Because $x_1^2 x_2^6 x_3$ is common to the evaluated monomial and all its derivatives, we call $x_1^2 x_2^6 x_3$ a *common factor*. This section is concerned with the evaluation of all common factors.

The kernel to compute common factors operates in two stages:

1. each of the first n threads of a thread block computes sequentially powers from the 2nd to the $(d-1)$ th of one of the n variables;
2. each of the threads of a block computes a common factor for one of the monomials of the system, as a product of k quantities computed at the first stage of the kernel.

Storing the values of the successive variables of the system in the successive locations of the global memory enables their coalesced reading into the shared memory by the threads of a warp, thus providing a fast input for the first stage of the kernel.

Both stages of the kernel are largely SIMT (Single Instruction Multiple Thread) routines since at the first stage each busy thread performs the same, $d - 1$, number of multiplications, and at the second stage each thread in each warp also performs the same, $k - 1$, number of multiplications.

The precomputed powers of variables are stored at the shared memory of the blocks, since these powers essentially constitute shared input data for the threads of the block while the threads are working on the second stage of the kernel. The powers are stored in shared memory in a two dimensional array `POWERS` of complex numbers, where the (i, j) th element represents the i th power of the j th variable. Such indexing is aimed at minimizing the number of shared memory bank conflicts at least during the first stage of the kernel, as different threads in a warp, after computing the current power of the variables designated to them, will be writing the power values into different banks of the shared memory.

As the threads of a block perform the second stage of the kernel, each thread computes a product of k quantities. These k quantities were computed at the first stage. As a thread proceeds to the next element in a product, to know what element to access in the shared memory array `POWERS`, it needs to know which variable and what exponent appears next in the monomial it is computing. The information about positions of variables and their exponents does not change during path tracking and is thus stored in the constant memory of the card. We reserve two arrays of unsigned chars `POSITIONS` and `EXPONENTS` in the constant memory to represent this information. Each element in `POSITIONS` represents a position of a variable from 0 to 255 in one of monomials of the system, and the element with the same index in `EXPONENTS` represents the degree of this variable decreased by one in the same monomial, giving us opportunity to work with variables appearing in degrees up to 255.

We need at least about 1,000 monomials to keep all 14 multiprocessors of our card well occupied for the algorithms we consider here, so several warps work on each multiprocessor simultaneously to hide long latency operations. This and the capacity of the constant memory, 65,536 bytes, prescribes working dimensions for our polynomial evaluation. Those dimensions are ranging from 30 to 40, if we want to keep m , the number of monomials in the polynomial, to be equal roughly to the dimension of the system, and k , the number of variables in the monomial, about half of the dimension. Indeed: for dimension 30 we would have 900 monomials, with a need of $900 \times 2 \times 15 \leq 30,000$ bytes; for dimension 40 we would have 1,600 monomials, with a need of $1,600 \times 2 \times 20 = 64,000$ bytes.

We close this subsection with a discussion about extensions and alternatives to this first kernel.

We are planning to introduce more compact encodings for storing the positions and exponents of the variables in the constant memory so to be working with higher dimensions. The more compact encodings might introduce some branching for the threads of a warp, after the decoded indexing information would be read from the constant memory into the registers of the block, while each thread in a warp would be encoding the actual position and exponent of the next variable power, which it needs to use for its computations. However the computations, which would follow encodings (the multiplications), where the threads of a warp will join again one path of execution, are supposed to dominate encodings in time, especially if higher precision

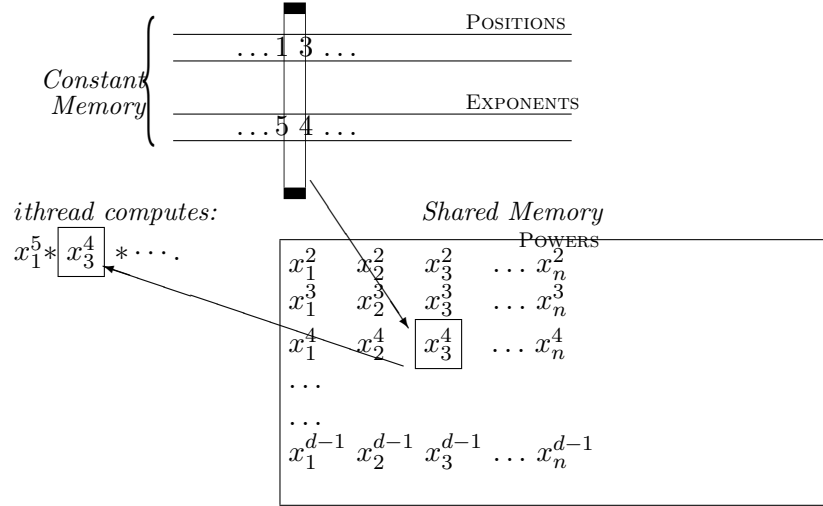


Figure 1: During the second stage of the first kernel, as the i th thread of the grid computes the common factor of the i th monomial of the system, it retrieves from the constant memory the variable position and exponent of each next variable power in the factor. Thus it knows which entry from the two dimensional array **POWERS** in shared memory to use next in the multiplication. The powers of the variables are computed during the first stage of the kernel by the threads of the block.

multiplications would be used. Thus with new ways of decoding, incorporated to store more efficiently monomial information in the constant memory, and employed multiprecision, we hope increase working dimensions for our implementation.

After each thread of a block computes its common factor, the successive threads of the block conveniently write their output values (one value per thread) into successive locations of the global memory, thus providing a coalesced output for the kernel.

As an alternative to computing common factors with the two above stages, one can skip precomputing powers, and assign to each thread all work, which is necessary for computing of assigned to it common factor, to do by itself from scratch. This could be done entirely in registers assigned to a block, without any use of shared memory. However this would introduce branching in execution of threads of a warp when monomials would have different tuples of exponents, and if one would choose that each thread would compute all powers up to $d - 1$ for participating in its monomial variables, it would most likely cause extensive multiple exponentiation of the same variables by threads within warps since the same variables tend to appear in multiple monomials of a system.

In our algorithm powers of variables are also computed multiple times – each block of threads computes its own copy of the set of powers from 2 to $d - 1$ for all n variables of the system. This might look as a drawback of the algorithm. However, for our working dimensions ranging from 30 to 40, and the number of maximal cores for one multiprocessor, 32, we would need to assign at most two blocks to work on precomputing degrees if we want to do it only once, in this case 12 of 14 multiprocessors would be idle during precomputing powers of variables. Also to start using the other 12 multiprocessors for the second stage of computing common factors, we would need to write the precomputed powers into the global memory, then to invoke a separate kernel with enough blocks to occupy all multiprocessors, and then threads of each block of the new kernel will access the global memory again for reading the powers of variables stored there. Our algorithm, as an alternative to prompted by the just described two kernels scheme additional time cost for global memory reading and writing, introduces the additional time cost, which is illustrated well by the following example.

Consider a system of dimension 32 with 28 monomials in each polynomial. If we will work with blocks of 32 threads, 28 blocks of threads will be launched. Then, in the worst case, if only one block will be occupying one multiprocessor at a time, the execution time for our two-stages kernel will be the same as if one block of 32 threads would be launched two times in a row to compute altogether 64 common factors. In particular, precomputing powers, despite in fact it would be done 28 times, time-wise would take the same amount of time as it would be done twice. Then, as within one thread block powers of all variables are computed in parallel, for our example then precomputing degrees would take in the worst case the same time as is needed for one core to compute $2(d - 2)$ multiplications (variables for the common factors need to be raised up to the power $d - 1$, which requires $d - 2$ multiplications). The degree d is in most cases not that high (while still allowing high total degrees for monomials). Thus multiple precomputing powers of variables in our two stages one kernel algorithm in most cases would compensate for the additional necessary global memory accesses as the powers are precomputed only once, and most likely, even reduce the computational time for precomputing powers.

3.2 Monomial Evaluation and Differentiation of Products of Variables

In this section we describe the implementation of the algorithm to evaluate a product of variables $x_{i_1}x_{i_2}\cdots x_{i_k}$ and all its derivatives.

We use the following terminology. For a term $\beta = c_a x_{i_1}^{a_{i_1}} x_{i_2}^{a_{i_2}} \cdots x_{i_k}^{a_{i_k}}$ in a polynomial f of the system, we refer to $\gamma := x_{i_1}^{a_{i_1}} x_{i_2}^{a_{i_2}} \cdots x_{i_k}^{a_{i_k}}$ as the corresponding monomial. Also, for convenience, we refer to monomials $x_{i_1}^{a_{i_1}-1} x_{i_2}^{a_{i_2}} \cdots x_{i_k}^{a_{i_k}}$, \dots , $x_{i_1}^{a_{i_1}} x_{i_2}^{a_{i_2}} \cdots x_{i_k}^{a_{i_k}-1}$ as to monomial derivatives of γ with respect to x_{i_1}, \dots, x_{i_k} correspondingly, despite they differ from the actual partial derivatives of γ by constant factors. Note that the coefficients $(c_a a_{i_1}), \dots, (c_a a_{i_k})$ of the corresponding to these monomials terms $\frac{\partial \beta}{\partial x_{i_1}}, \dots, \frac{\partial \beta}{\partial x_{i_k}}$ in $\frac{\partial f}{\partial x_{i_1}}, \dots, \frac{\partial f}{\partial x_{i_k}}$ are precomputed before the actual path tracking begins. Finally we refer to f as to a hosting polynomial of the term β (and of the monomial γ) in the system, and we refer to $\frac{\partial f}{\partial x_{i_1}}, \dots, \frac{\partial f}{\partial x_{i_k}}$ as to hosting polynomials of the terms $\frac{\partial \beta}{\partial x_{i_1}}, \dots, \frac{\partial \beta}{\partial x_{i_k}}$ (and of the corresponding monomial derivatives of γ) in the Jacobian of the system.

In our second kernel each thread first computes one monomial and its monomial derivatives. Secondly it multiplies the computed value of the monomial by its coefficient in the hosting that monomial polynomial of the system, as well as it multiplies the values of the computed derivatives of the monomial by their coefficients in the hosting those monomial derivatives polynomials of the Jacobian. Thus this kernel completes computing additive terms of the polynomials of the system and the Jacobian, and the third last kernel only adds appearing in each polynomial terms to finish evaluating polynomials (of the system and of the Jacobian).

A thread of the second kernel performs only $5k - 4$ multiplications and uses $k + 1$ complex double locations of shared memory L_1, L_2, \dots, L_{k+1} and one variable in registers to perform all the announced above work. As was discussed in the previous section, through an example, we obtain the monomial derivatives of a monomial $\mathbf{x}^{\mathbf{a}}$ by multiplying the common factor $x_{i_1}^{a_{i_1}-1} x_{i_2}^{a_{i_2}-1} \cdots x_{i_k}^{a_{i_k}-1}$ by the partial derivatives of $x_{i_1}x_{i_2}\cdots x_{i_k}$. It takes $3k - 6$ multiplications out of $5k - 4$ multiplications performed by a thread to compute partial derivatives of a Speelpenning product. Another k multiplications are performed to multiply the common factor by the values of derivatives of a Speelpenning product to obtain monomial derivatives. One additional multiplication is done to obtain the value of the monomial itself as a product of a monomial derivative with respect to any of participating in the monomial variables and the value of that variable. Finally a thread performs another $k + 1$ multiplications to multiply the values of the monomial and its monomial derivatives by the coefficients. The work of a thread of the second kernel is illustrated in Table 1.

To obtain derivatives of a Speelpenning product a thread first stores x_{i_1} in the location L_2 . Then it computes sequentially, by $k - 2$ multiplications, the $k - 2$ forward products $x_{i_1}x_{i_2}$, $x_{i_1}x_{i_2}x_{i_3}$, \dots , $x_{i_1}x_{i_2}x_{i_3}\cdots x_{i_{k-1}}$, for each new r ranging from 1 to $k - 2$ obtaining the product $x_{i_1}x_{i_2}x_{i_3}\cdots x_{i_{r+1}}$ as $(x_{i_1}x_{i_2}x_{i_3}\cdots x_{i_r})x_{i_{r+1}}$ and storing the newly obtained forward product into location L_{r+2} . Eventually the locations L_3, \dots, L_k are filled with the $k - 2$ obtained forward products. Note that at this point the location L_k contains the derivative of the Speelpenning product with respect to x_{i_k} . In registers of the block we keep the only complex double variable

L_1	L_2	L_3	L_4	Q
	x_1			
	x_1	$x_1 * x_2$		
	x_1	$x_1 x_2$	$(x_1 x_2) * x_3$	
	x_1	$(x_1 x_2) * x_4$	$x_1 x_2 x_3$	x_4
	$x_1 * (x_3 x_4)$	$x_1 x_2 x_4$	$x_1 x_2 x_3$	$x_4 * x_3$
$x_2 x_3 x_4$	$x_1 x_3 x_4$	$x_1 x_2 x_4$	$x_1 x_2 x_3$	$(x_4 x_3) * x_2$
$\frac{\partial s}{\partial x_1}$	$\frac{\partial s}{\partial x_2}$	$\frac{\partial s}{\partial x_3}$	$\frac{\partial s}{\partial x_4}$	

L_1	L_2	L_3	L_4	L_5
$\frac{\partial s}{\partial x_1} * \alpha$	$\frac{\partial s}{\partial x_2} * \alpha$	$\frac{\partial s}{\partial x_3} * \alpha$	$\frac{\partial s}{\partial x_4} * \alpha$	
$\frac{1}{3} \frac{\partial \gamma}{\partial x_1}$	$\frac{1}{7} \frac{\partial \gamma}{\partial x_2}$	$\frac{1}{4} \frac{\partial \gamma}{\partial x_3}$	$\frac{1}{5} \frac{\partial \gamma}{\partial x_4}$	
$\frac{1}{3} \frac{\partial \gamma}{\partial x_1}$	$\frac{1}{7} \frac{\partial \gamma}{\partial x_2}$	$\frac{1}{4} \frac{\partial \gamma}{\partial x_3}$	$\frac{1}{5} \frac{\partial \gamma}{\partial x_4}$	$\frac{1}{5} \frac{\partial \gamma}{\partial x_4} * x_4$
$\frac{1}{3} \frac{\partial \gamma}{\partial x_1}$	$\frac{1}{7} \frac{\partial \gamma}{\partial x_2}$	$\frac{1}{4} \frac{\partial \gamma}{\partial x_3}$	$\frac{1}{5} \frac{\partial \gamma}{\partial x_4}$	γ
$\frac{1}{3} \frac{\partial \gamma}{\partial x_1} * (3c)$	$\frac{1}{7} \frac{\partial \gamma}{\partial x_2} * (7c)$	$\frac{1}{4} \frac{\partial \gamma}{\partial x_3} * (4c)$	$\frac{1}{5} \frac{\partial \gamma}{\partial x_4} * (5c)$	$\gamma * c$
$\frac{\partial \beta}{\partial x_1}$	$\frac{\partial \beta}{\partial x_2}$	$\frac{\partial \beta}{\partial x_3}$	$\frac{\partial \beta}{\partial x_4}$	β

Table 1: Example: A thread of the second kernel computes the term $\beta = c x_1^3 x_2^7 x_3^4 x_4^5$ and its derivatives by $5k - 4 = 5 * 4 - 4 = 16$ multiplications, using $k + 1 = 5$ shared memory locations and one local variable. Here $\gamma := \frac{1}{c} \beta = x_1^3 x_2^7 x_3^4 x_4^5$ is the corresponding monomial, $\alpha := x_1^2 x_2^6 x_3^3 x_4^4$ (the common factor), and $s := x_1 x_2 x_3 x_4$ (the Speelpenning product). Note that the coefficients $(3c), (7c), (4c), (5c)$ are precomputed, only explicit $*$ -s stay for performed multiplications.

Q to store the current backward product. We initialize Q with x_{i_k} . A thread computes the derivative of the Speelpenning product with respect to $x_{i_{k-1}}$ at L_{k-1} by multiplying stored in that location the forward product $x_{i_1} x_{i_2} x_{i_3} \cdots x_{i_{k-2}}$ by the current value of Q , which is x_{i_k} .

In the next $k - 3$ steps, each of which consists of two multiplications, we compute partial derivatives of the Speelpenning product with respect to $x_{i_2}, x_{i_3}, \dots, x_{i_{k-2}}$, and store the computed values in locations L_2, L_3, \dots, L_{k-2} . At the r th step, as r ranges from 1 to $k - 3$, the Q represents the backward product $x_{i_k} x_{i_{k-1}} \cdots x_{i_{k-r}}$. At the r th step we first update the value of Q , accordingly to its above definition, by one multiplication as $Q = Q \times x_{i_{k-r}}$. The second multiplication updates the shared memory location L_{k-r-1} as $L_{k-r-1} = L_{k-r-1} \times Q$, so to obtain in this location the partial derivative of Speelpenning product with respect to $x_{i_{k-r-1}}$ as a product of previously stored there forward product $x_{i_1} x_{i_2} x_{i_3} \cdots x_{i_{k-r-2}}$ times the current backward product $x_{i_k} x_{i_{k-1}} \cdots x_{i_{k-r}}$.

Finally we obtain the last yet not obtained partial derivative of Speelpenning product with respect to x_{i_1} at Q , by the product $Q = Q \times x_{i_2}$ and store the obtained value at the shared memory location L_1 .

The above procedure prescribes to a thread to perform $k - 2$ multiplications to obtain the forward products, $k - 2$ multiplications to obtain the backward products, and $k - 2$ multiplications of backward and forward products. Thus indeed, the total number of multiplications for obtaining all derivatives of the Speelpenning product equals $3k - 6$.

Now a thread computes monomial derivatives in locations L_1, L_2, \dots, L_k by multiplying stored

in these locations values of derivatives of a Speelpenning product by the common factor computed in the first kernel. Then it computes the value of the monomial itself as the product of its monomial derivative with respect to x_{i_k} , stored in L_k times the value of x_{i_k} . It stores the computed monomial value at L_{k+1} . Finally it multiplies each of the values stored in L_1, L_2, \dots, L_{k+1} , i.e.: the values of the monomial and its monomial derivatives, by the corresponding coefficients.

As we take the same k — the number of variables in a monomial — for all monomials of the system, each thread of the second kernel will go through the same path of execution for the entire list of instructions of the kernel, which largely amounts to $5k - 4$ complex double multiplications. Thus all 32 threads within each warp will be indeed doing all the prescribed work for the assigned to them 32 monomials in a parallel fashion on an available multiprocessor.

We close this subsection with some memory considerations. Let B denote the number of threads in a block.

In addition to the fast access space of $B(k + 1)$ locations equally divided between threads of a block for storing their intermediate results, as the threads proceed along the kernel, we reserve in shared memory of a block a space for values of all variables of the system. The values of the variables are subject shared use of the threads of each block, as the same variables appear in different monomials. Thus, provided values of successive variables are stored in successive locations of global memory, and working with $n = 32$, $k = 16$, $B = 32$, we would need to access global memory only once by all threads of a block simultaneously, as each thread would request a value of one variable, to download the values of all 32 variables into the shared memory of the block for their further common use by all threads of the block.

At the same time, if shared memory would not be used for storing values of variables, each thread would need to access global memory at least 16 times to get the values of all appearing in its monomial variables. The shared memory capacity allows us to apply the above algorithm of the second kernel for our working dimensions ranging between 30 and 40 as well as for some larger dimensions. We also could increase precision from double to double double and still work with dimensions up to 70, as long as k is less or equal than a half of dimension. Indeed, each thread would need for treating its monomial $k + 1$ complex double double locations, thus

$$\begin{aligned} & (n/2 + 1) \times 2 \times \text{sizeof}(\text{double double}) \\ & \leq (70/2 + 1) \times 2 \times 16 = 1,152 \end{aligned}$$

bytes in shared memory. To treat 32 monomials by a block of 32 threads we would need then at most $32 \times 1,152 = 36,864$ bytes of shared memory. Adding to this

$$\begin{aligned} & n \times \text{sizeof}(\text{complex double double}) \\ & \leq 70 \times 2 \times \text{sizeof}(\text{double double}) \\ & = 70 \times 2 \times 16 = 2,240 \end{aligned}$$

bytes in shared memory for storing values of the variables, we are still $(49,152 - (36,864 + 2,240)) > 10,000$ bytes below the capacity of the shared memory of a block.

Another important note about the memory management is that the array `POSITIONS` in constant memory, which contains positions indexes of variables in the monomials, and used in the first kernel, is used in this kernel as well, as threads are determining what variable in the

shared memory to access as they need to perform each new multiplication while updating their forward and backward products.

3.3 Multiplication of Evaluated Monomials with Coefficients and Summation of Terms

In the third stage the evaluated monomials first are multiplied with their coefficients in the polynomials of the system or the Jacobian.

The coefficients are stored in the global memory, since the capacity of the constant memory is exhausted by the variables positions indexes and variables exponents information. As we multiply monomials and their derivatives by the coefficients, we need to read the values of coefficients from the global memory fast. The total number of monomials in the system is $n \times m$. For mapping purposes all the monomials are ordered in a sequence S_m of length $n \times m$. For instance the monomials in S_m might be ordered as following: first m elements of the sequence are the monomials of the first polynomial, the next m elements are the monomials of the second polynomial, and so on. The coefficients are stored during entire path tracking in an array COEFFS of length $n \times m \times (k + 1)$, which is the total number of monomials in the system and its Jacobian. The coefficients in COEFFS are stored in the following order:

- The first element of COEFFS is the coefficient of the derivative of the first monomial in S_m with respect to its first variable;
- the second element of Coeffs is the coefficient of the derivative of the second monomial in S_m with respect to its first variable, and so on until
- the nm th element of COEFFS, which is the coefficient of the derivative of the last monomial in S_m with respect to its first variable.
- The next $n \times m$ elements of COEFFS are the coefficients of the derivatives of monomials from S_m , with respect to the monomials second variables, also listed in accordance with order in S_m .

The portions of nm coefficients come in a similar manner until the k th portion of nm coefficients, in which are stored, in order inherited from S_m , the coefficients of monomial derivatives with respect to the monomials k th (last) variables. The last $(k + 1)$ th portion of the nm coefficients contains actually the coefficients of the system in order prescribed by order in S_m . With this way of storing coefficients, if i th thread of the second kernel is in charge of i th monomial in S_m for each $i = 1, 2, \dots, nm$, we largely obtain a coalesced access within warps, as threads of a warp prescribed simultaneously to access the coefficients of their monomials or the j th, $j \in \{1, 2, \dots, k\}$, derivative's coefficients of their monomials in COEFFS.

After multiplying the monomial and their derivatives values by coefficients, which is the last computational step of the second kernel, it is just left to add the corresponding computed additive terms to obtain the values of the polynomials of the system and of the Jacobian. If the size of a thread block used for the execution of the second kernel is smaller than m , then monomials of each polynomial of the system are treated by multiple blocks of the second kernel. In this case, even if some of the involved summations are done yet by the threads of the second kernel,

it is necessary to launch another kernel to combine partial sums which are obtained by different blocks of the second kernel, which are working on monomials of the same polynomials. The situation, when the size of a thread block of the second kernel is less than m , is very common for our working dimensions: we try to keep the block size of the second kernel equal to 32, because of described above shared memory limited capacity considerations, on the other hand, we are willing to work with higher dimensions, ranging from 50 to 70, while we want to keep $m \approx n$. Also, computing partial sums for polynomials of the Jacobian by threads of the second kernel would involve branching in execution paths of the threads within warps, as different subsets of variables appear in monomials treated by different threads within a warp. Because of the above reasons we decided to introduce a third kernel, which would perform all involved summations, so to complete obtaining the values of the polynomials, as all multiplicative operations are done by the first two kernels.

Each thread of the third kernel sums additive terms of one of $n^2 + n$ polynomials of the combined set of polynomials of the system and the Jacobian matrix. To make each thread to go through the same execution path, all what we assign to each thread to do during the execution of the kernel is to add exactly m terms. Thus, if a thread computes the value of the derivative of the p th polynomial with respect to x_i , and a j th monomial in the p th polynomial does not contain x_i , the thread which computes the derivative of the p th polynomial with respect to x_i , at the j th step does add to its current partial sum zero – the zero monomial derivative, which we probably never would add in a CPU execution. To ensure this, without introducing any if statements, the output array of the second kernel in the global memory along with its meaningful $nm(k+1)$ locations (the number of monomials and monomial derivatives of the system) contains also $(n^2 + n)m - nm(k+1)$ locations, the values at which are originally set and kept to store zero values along the entire path tracking. These zero locations represent the zero monomial derivatives as in the described above situation. We also wish that the threads within warps of the third kernel for each step j , $j = 1, 2 \dots m$ would perform a coalesced reading of the input data entries. To allow coalesced reading of the values of monomials and their derivatives by the threads of the third kernel, and to introduce the $(n^2 + n)m - nm(k+1)$ zero monomial derivatives, the output of the second kernel is stored in the global memory in array MONS in the format we explain next.

The size of the array MONS is $(n^2 + n)m$, representing the terms in $n^2 + n$ summations, m terms each. The first $n^2 + n$ elements of the array represent the first terms in each of $n^2 + n$ summations (polynomials). In particular, these first $n^2 + n$ elements are: the first n elements are the first monomials of the polynomials of the system, the second n elements are the derivatives of the first monomials with respect to x_1 , the third n elements are the derivatives of the first monomials with respect to x_2 , and so on until the $(n+1)$ th n elements, which are the derivatives of the first monomials with respect to x_n . The second $n^2 + n$ elements represent the second terms in each of $n^2 + n$ summations, and again the first n elements of them represent the second monomials of the polynomials of the system, and the next n^2 elements represent the partial derivatives of the second monomials of the system, listed in the same order as are listed the derivatives of the first monomials. In general the j th $n^2 + n$ elements represent j th monomials of the polynomials of the system and their partial derivatives listed in the same order as listed the first monomials of the system and their partial derivatives at the first $n^2 + n$ elements of the array.

For simplicity in this description we assumed that the number B of threads in a block, the block

size, divides n^2+n . Now if we launch $(n^2+n)/B$ blocks, with a thread $t = BlockId \times B + ThreadId$ computing the sum: $\sum_{j=0}^{m-1} MONS[t + j(n^2 + n)]$, the obtained sums will represent the values of polynomials of the system and of the Jacobian, while access to the elements of MONS will be coalesced within warps at each step $j = 0, 1, \dots, m-1$ of the summation. To create the array MONS in such a format, we had to make the threads of the second kernel to output the values of monomials and their derivatives not in a coalesced way. However there was a tradeoff:

- either to make the output of the second kernel coalesced and then the input of the third kernel could not be accessed in a coalesced way,
- or as we chose to provide ability for the threads of the third kernel to read the input data in a coalesced way, and paid the price of not coalesced writing of the output of the second kernel.

4 Computational Experiments

Our computations are done on a HP Z800 workstation, running Red Hat Enterprise Linux Workstation release 6.1. The CPU is an Intel Xeon X5690 at 3.47 Ghz. The processor clock of the NVIDIA Tesla C2050 Computing Processor runs at 1147 Mhz. The graphics card has 14 multiprocessors, each with 32 cores, for a total of 448 cores. As the clock speed of the GPU is a third of the clock speed of the CPU, we hope to achieve a double digit speedup. We used the NVIDIA CUDA compiler driver `nvcc`, release 4.0, V0.2.1221.

In Table 2 and 3 we list results of our preliminary implementation. The number of threads in each block was 32 for all three kernels to evaluate a system and its Jacobian matrix of dimension 32. Generating 32 monomials per polynomial leads to 1,024 monomial in total.

#monomials	Tesla C2050	1 CPU core	speedup
704	14.514 sec	1min 50.9 sec	7.60
1024	15.265 sec	2min 39.3 sec	10.44
1536	17.000 sec	3min 58.7 sec	14.04

Table 2: Wall clock times and speedups for 100,000 evaluations of a polynomial system and its Jacobian matrix of dimension 32. Each monomial has 9 variables occurring with nonzero power of at most 2.

#monomials	Tesla C2050	1 CPU core	speedup
704	19.068 sec	3min 16.9 sec	10.33
1024	20.800 sec	4min 43.3 sec	13.62
1536	21.763 sec	7min 05.8 sec	19.56

Table 3: Wall clock times and speedups for 100,000 evaluations of a polynomial system and its Jacobian matrix of dimension 32. Each monomial has 16 variables occurring with nonzero power of at most 10.

Increasing the number of monomials to 2,048 in Table 2 and 3 would have yielded a speedup of more than 20, but the capacity of the constant memory was not sufficient to hold the exponents and positions of all 2,048 monomials. For larger systems, we will upgrade our preliminary implementation with a better compression strategy (instead of the current `char` used for each exponent).

5 Conclusions

In this paper we describe a first implementation on the NVIDIA Tesla C2050 Computing Processor using the CUDA compiler tools to evaluate polynomials in several variables and all their derivatives. Our implementation consists of three kernels. The first kernel computes factors common to all monomial derivatives, factors that are multiplied with the Speelpenning products and their derivatives, computed by the second kernel. The third kernel adds the evaluated monomials and derivatives and returns the values of the polynomial system and its Jacobian matrix.

We obtained good speedups for randomly generated polynomial systems of dimension 32 (the warp size) and fixed number of monomials per polynomial. For a double digit speedup, we need to have at least 1,000 monomials. The speedups improve for higher degree monomials.

References

- [1] S.G. Akl. Superlinear performance in real-time parallel computation. *The Journal of Supercomputing*, 29(1):89–111, 2004.
- [2] E.L. Allgower and K. Georg. *Introduction to Numerical Continuation Methods*, volume 45 of *Classics in Applied Mathematics*. SIAM, 2003.
- [3] D.C.S. Allison, A. Chakraborty, and L.T. Watson. Granularity issues for solving polynomial systems via globally convergent algorithms on a hypercube. *J. of Supercomputing*, 3(1):5–20, 1989.
- [4] D.J. Bates, J.D. Hauenstein, and A.J. Sommese. A parallel endgame. In L. Gurvits, P. Pébay, J.M. Rojas, and D. Thompson, editors, *Randomization, Relaxation, and Complexity in Polynomial Equation Solving*, volume 556 of *Contemporary Mathematics*, pages 25–35. AMS, 2011.
- [5] C. Bischof, N. Guertler, A. Kowartz, and A. Walther. Parallel reverse mode automatic differentiation for OpenMP programs with ADOL-C. In C. Bischof, H.M. Bücker, P. Hovland, U. Naumann, and J. Utke, editors, *Advances in Automatic Differentiation*, pages 163–173. Springer-Verlag, 2008.
- [6] A. Chakraborty, D.C.S. Allison, C.J. Ribbens, and L.T. Watson. The parallel complexity of embedding algorithms for the solution of systems of nonlinear equations. *IEEE Transactions on Parallel and Distributed Systems*, 4(4):458–465, 1993.
- [7] T.J. Dekker. A floating-point technique for extending the available precision. *Numerische Mathematik*, 18(3):224–242, 1971.

- [8] P. Emeliyanenko. Efficient multiplication of polynomials on graphics hardware. In Y. Dou, R. Gruber, and J.M. Joller, editors, *Advanced Parallel Processing Technologies. 8th International Symposium, APPT 2009, Rapperswil, Switzerland, August 2009*, volume 5737 of *Lecture Notes in Computer Science*, pages 134–149. Springer-Verlag, 2009.
- [9] P. Emeliyanenko. A complete modular resultant algorithm targeted for realization on graphics hardware. In M.M. Maza and J.-L. Roch, editors, *Proceedings of the 4th International Workshop on Parallel Symbolic Computation (PASCO 2010), July 21-23 2010, Grenoble, France*, pages 35–43. ACM, 2010.
- [10] P. Emeliyanenko. High-performance polynomial GCD computations on graphics processors. In W.W. Smari and J.P. McIntire, editors, *Proceedings of the 2011 International Conference on High Performance Computing & Simulation (HPCS 2011)*, pages 215–224. IEEE, 2011.
- [11] M. Grabner, T. Pock, T. Gross, and B. Kainz. Automatic differentiation for GPU-accelerated 2D/3D registration. In C. Bischof, H.M. Bücker, P. Hovland, U. Naumann, and J. Utke, editors, *Advances in Automatic Differentiation*, pages 259–269. Springer-Verlag, 2008.
- [12] A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM, second edition, 2008.
- [13] Y. Guan and J. Verschelde. Parallel implementation of a subsystem-by-subsystem solver. In *The proceedings of the 22th High Performance Computing Symposium, Quebec City, 9-11 June 2008*, pages 117–123. IEEE Computer Society, 2008.
- [14] T. Gunji, S. Kim, K. Fujisawa, and M. Kojima. PHoMpara – parallel implementation of the Polyhedral Homotopy continuation Method for polynomial systems. *Computing*, 77(4):387–411, 2006.
- [15] Y. Hida, X.S. Li, and D.H. Bailey. Algorithms for quad-double precision floating point arithmetic. In *15th IEEE Symposium on Computer Arithmetic (Arith-15 2001), 11-17 June 2001, Vail, CO, USA*, pages 155–162. IEEE Computer Society, 2001. Shortened version of Technical Report LBNL-46996, software at <http://crd.lbl.gov/~dhbailey/mpdist/qd-2.3.9.tar.gz>.
- [16] W.W. Hwu (editor). *GPU Computing Gems: Emerald Edition*. Morgan Kaufmann, 2011.
- [17] D.B. Kirk and W.W. Hwu. *Programming Massively Parallel Processors. A Hands-on Approach*. Morgan Kaufmann, 2010.
- [18] M. Kojima. Efficient evaluation of polynomials and their partial derivatives in homotopy continuation methods. *Journal of the Operations Research Society of Japan*, 51(1):29–54, 2008.
- [19] A. Leykin. Numerical algebraic geometry. *The Journal of Software for Algebra and Geometry: Macaulay2*, 3:5–10, 2011.
- [20] A. Leykin and J. Verschelde. Factoring solution sets of polynomial systems in parallel. In T. Skeie and C.-S. Yang, editors, *Proceedings of the 2005 International Conference on*

- Parallel Processing Workshops. 14-17 June 2005. Oslo, Norway. High Performance Scientific and Engineering Computing*, pages 173–180. IEEE Computer Society, 2005.
- [21] A. Leykin and J. Verschelde. Decomposing solution sets of polynomial systems: a new parallel monodromy breakup algorithm. *The International Journal of Computational Science and Engineering*, 4(2):94–101, 2009.
 - [22] A. Leykin, J. Verschelde, and Y. Zhuang. Parallel homotopy algorithms to solve polynomial systems. In N. Takayama and A. Iglesias, editors, *Proceedings of ICMS 2006*, volume 4151 of *Lecture Notes in Computer Science*, pages 225–234. Springer-Verlag, 2006.
 - [23] T.Y. Li. Numerical solution of polynomial systems by homotopy continuation methods. In F. Cucker, editor, *Handbook of Numerical Analysis. Volume XI. Special Volume: Foundations of Computational Mathematics*, pages 209–304. North-Holland, 2003.
 - [24] T.Y. Li and C.-H. Tsai. HOM4PS-2.0para: Parallelization of HOM4PS-2.0 for solving polynomial systems. *Parallel Computing*, 35(4):226–238, 2009.
 - [25] M. Lu, B. He, and Q. Luo. Supporting extended precision on graphics processors. In A. Ailamaki and P.A. Boncz, editors, *Proceedings of the Sixth International Workshop on Data Management on New Hardware (DaMoN 2010)*, June 7, 2010, Indianapolis, Indiana, pages 19–26, 2010. Software at <http://code.google.com/p/gpugrpec/>.
 - [26] M.M. Maza and W. Pan. Fast polynomial multiplication on a GPU. *Journal of Physics: Conference Series*, 256, 2010. High Performance Computing Symposium (HPCS2010), 5-9 June 2010, Victoria College, University of Toronto, Canada.
 - [27] M.M. Maza and W. Pan. Solving bivariate polynomial systems on a GPU. *Journal of Physics: Conference Series*, 341(1):1–19, 2012.
 - [28] A. Morgan. *Solving polynomial systems using continuation for engineering and scientific problems*. Prentice-Hall, 1987. Volume 57 of Classics in Applied Mathematics Series, SIAM 2009.
 - [29] NVIDIA. NVIDIA CUDA Programming Guide. Version 3.0. 2010.
 - [30] D.N. Priest. *On Properties of Floating Point Arithmetics: Numerical Stability and the Cost of Accurate Computations*. PhD thesis, University of California at Berkeley, 1992. <ftp://ftp.icsi.berkeley.edu/pub/theory/priest-thesis.ps.Z>.
 - [31] S.M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287449, 2010.
 - [32] J.R. Shewchuk. Adaptive precision floating-point arithmetic and fast robust geometric predicates. *Discrete Comput. Geom.*, 18(3):305–363, 1997.
 - [33] M. Snir, S. Otto, S. Huss-Lederman, D. Walker, and J. Dongarra. *MPI - The Complete Reference Volume 1, The MPI Core*. Massachusetts Institute of Technology, second edition, 1998.

- [34] A.J. Sommese, J. Verschelde, and C.W. Wampler. Introduction to numerical algebraic geometry. In A. Dickstein and I.Z. Emiris, editors, *Solving Polynomial Equations. Foundations, Algorithms and Applications*, volume 14 of *Algorithms and Computation in Mathematics*, pages 301–337. Springer-Verlag, 2005.
- [35] A.J. Sommese and C.W. Wampler. *The Numerical solution of systems of polynomials arising in engineering and science*. World Scientific, 2005.
- [36] H.-J. Su, J.M. McCarthy, M. Sosonkina, and L.T. Watson. Algorithm 857: POLSYS_GLP: A parallel general linear product homotopy code for solving polynomial systems of equations. *ACM Trans. Math. Softw.*, 32(4):561–579, 2006.
- [37] J. Verschelde. Algorithm 795: PHCpack: A general-purpose solver for polynomial systems by homotopy continuation. *ACM Trans. Math. Softw.*, 25(2):251–276, 1999. Software available at <http://www.math.uic.edu/~jan/download.html>.
- [38] J. Verschelde and Y. Wang. Computing feedback laws for linear systems with a parallel Pieri homotopy. In Y. Yang, editor, *Proceedings of the 2004 International Conference on Parallel Processing Workshops, 15-18 August 2004, Montreal, Quebec, Canada. High Performance Scientific and Engineering Computing*, pages 222–229. IEEE Computer Society, 2004.
- [39] J. Verschelde and G. Yoffe. Quality up in polynomial homotopy continuation by multi-threaded path tracking. Preprint [arXiv:1109.0545v1](https://arxiv.org/abs/1109.0545v1) [cs.DC] 2 Sep 2011.
- [40] J. Verschelde and G. Yoffe. Polynomial homotopies on multicore workstations. In M.M. Maza and J.-L. Roch, editors, *Proceedings of the 4th International Workshop on Parallel Symbolic Computation (PASCO 2010), July 21-23 2010, Grenoble, France*, pages 131–140. ACM, 2010.
- [41] J. Verschelde and Y. Zhuang. Parallel implementation of the polyhedral homotopy method. In T.M. Pinkston and F. Ozguner, editors, *Proceedings of the 2006 International Conference on Parallel Processing Workshops. 14-18 Augustus 2006. Columbus, Ohio. High Performance Scientific and Engineering Computing*, pages 481–488. IEEE Computer Society, 2006.
- [42] L.T. Watson. Probability-one homotopies in computational science. *J. Comput. Appl. Math.*, 140(1&2):785–807, 2002.