Optimal Operation of GaN Thin Film Epitaxy Employing Control Vector Parametrization

Amit Varshney and Antonios Armaou
Dept. of Chemical Engineering, Pennsylvania State University, University Park, PA 16802

DOI 10.1002/aic.10691
Published online November 23, 2005 in Wiley InterScience (www.interscience.wiley.com).

An approach that links nonlinear model reduction techniques with control vector parametrization-based schemes is presented, to efficiently solve dynamic constraint optimization problems arising in the context of spatially-distributed processes governed by highly-dissipative nonlinear partial-differential equations (PDEs), utilizing standard nonlinear programming techniques. The method of weighted residuals with empirical eigenfunctions (obtained via Karhunen-Loève expansion) as basis functions is employed for spatial discretization together with control vector parametrization formulation for temporal discretization. The stimulus for the earlier approach is provided by the presence of low order dominant dynamics in the case of highly dissipative parabolic PDEs. Spatial discretization based on these few dominant modes (which are elegantly captured by empirical eigenfunctions) takes into account the actual spatiotemporal behavior of the PDE which cannot be captured using finite difference or finite element techniques with a small number of discretization points/elements. The proposed approach is used to compute the optimal operating profile of a metallorganic vapor-phase epitaxy process for the production of GaN thin films, with the objective to minimize the spatial nonuniformity of the deposited film across the substrate surface by adequately manipulating the spatiotemporal concentration profiles of Ga and N precursors at the reactor inlet. It is demonstrated that the reduced order optimization problem thus formulated using the proposed approach for nonlinear order reduction results in considerable savings of computational resources and is simultaneously accurate. It is demonstrated that by optimally changing the precursor concentration across the reactor inlet it is possible to reduce the thickness nonuniformity of the deposited film from a nominal 33% to 3.1%.© 2005 American Institute of Chemical Engineers *AIChE J,* 52: 1378–1391, 2006

Keywords: dissipative partial differential equations, Karhunen-Loève expansion, spatially-distributed processes, GaN, metallorganic vapor phase epitaxy, dynamic optimization, control vector parametrization

Introduction

Most of the processes relevant to the chemical process industry necessitate the consideration of transport phenomena (fluid flow, heat, and mass transfer) often coupled with chemical reactions. Examples range from reactive distillation in petroleum processing to plasma enhanced chemical vapor deposition, etching and metallorganic vapor phase epitaxy (MOVPE) in semiconductor manufacturing. Mathematical descriptions of these transport-reaction processes can be derived from dynamic conservation equations and usually involve highly dissipative (typically parabolic) partial-differential equations (PDEs). Optimal operation of these processes with respect to certain economic criteria has always been an indus-

Correspondence concerning this article should be addressed to A. Armaou at armaou@psu.edu.

trial priority and a subject of intense research. Traditionally, this problem has been addressed assuming steady-state operating conditions since limited computational resources prevented expensive dynamic optimization frameworks. However, the ever increasing computational power has made it possible to take advantage of the transient evolution of processes towards achieving improved economic objectives.

The presence of constraints in the form of PDEs hinder the direct applicability of standard search algorithms for optimization with the exception of calculus of variations. Traditional practice to deal with such constraints is to discretize the PDEs in both spatial and temporal coordinates using finite differences or finite elements or collocation techniques, and apply techniques for large scale sparse nonlinear programs (NLPs) to the resulting optimization problem, such as reduced gradient and reduced successive quadratic programming (see for example,[1,2,3]). The size of the NLP is dictated by the number of gridpoints used for discretization which can become prohibitive (from a computational point of view) if one desires to accurately capture the spatiotemporal behavior of the PDE. To overcome this issue, nonlinear model reduction techniques for spatially distributed transport-reaction processes operating at steady state[4,5] and unsteady-state[6] were recently employed to formulate approximate low-order optimization problems. In these investigations spatial discretization was carried out using the method of weighted residuals, using empirical eigenfunctions as basis functions. These eigenfunctions were generated by the application of Karhunen-Loève expansion (also known as Proper Orthogonal Decomposition, Principle Component Analysis and Method of Empirical Eigenfunctions[7,8,9]) on an ensemble of solution data of the PDEs for the span of process parameters. The motivation behind this approach was the presence of finite number of dominant spatial patterns (eigenmodes) in the solution of highly dissipative PDEs which govern its long time dynamics, while the remaining infinite dimensional (stable) fast modes relax to these finite dimensional slow dynamics.[10,11] The NLPs thus formed are significantly lower in size, and can be solved faster than those that result from finite differences/elements. The principle reason that allows model reduction is that the spatiotemporal behavior of the given PDE system is accounted for in the shape of the empirical eigenfunctions. An additional reduction in the size of the NLP can be made if discretization is performed *only* for the vector of control variables, and ODE equality constraints are directly integrated in time (the control vector parametrization (CVP) scheme[12,13,14,15,16]). The advantage of the control vector parametrization scheme is that the optimization is performed for the reduced set of discretized decision variables rather than for the complete set of discretized variables. Incorporation of second-order derivative information into CVP framework for improved efficiency has also been addressed.[17,18] However, only a few studies have investigated the construction of low-order nonlinear programs using empirical eigenfunctions for spatial discretization and control vector parametrization for temporal discretization.[19,20,21]

There have been a few studies in the past that address the issues of optimization and control of thin film growth processes relevant to microelectronic industry. They include feedback control of plasma etching process,[22,23] order reduction and subsequent optimization and real time control of rapid thermal chemical vapor deposition,[24,25,26] and steady-state optimization of metal organic vapor phase epitaxy process.[27] However, the problem of optimal operation of MOVPE using a dynamic process model has not been addressed. MOVPE is a method of choice to produce a variety of high-performance optical and electronic devices including light-emitting diodes, quantum-well lasers, and heterojunction bipolar transistors. Multilayered structures of group-III nitrides form the basis of these devices. Among these, GaN and related semiconductors (for example AlGaN, AlGaInN, GaInN) have been the focus of intense industrial and academic research since the last decade. Their large direct bandgap energy (for example 3.4 eV in case of GaN) makes them promising semiconductors in the manufacture of blue-green LEDs and laser diodes. MOVPE utilizes the thermal decomposition and reaction of gaseous precursors to epitaxially grow multiple layers of III-nitride thin films with precise thickness, composition and dopant level. The success of the deposition process and the quality of the devices depend heavily on the film thickness (which is in the order of a few Å), and the sharpness of the composition profile at the heterostructure interface.[28] In order to grow films of uniform thickness, a spatially invariant concentration of Ga containing precursors over the deposition-substrate is necessary; a requirement which is impossible to meet because of transport and reaction limitations. Recent simulation results[29] show that the thickness nonuniformity is approximately 25%. Thickness uniformity can, in principle, be improved by using multiple inlets and feeding precursors from alternate inlets into the reactor. However, such an implementation suffers the drawback of increased complexity, and increasing the number of inlets does not guarantee a high degree of thickness uniformity.

In this work, we present an approach that links nonlinear model reduction techniques with CVP-based formulation schemes to efficiently solve *dynamic* constraint optimization problems arising in the context of spatially-distributed processes governed by highly-dissipative nonlinear PDEs. The approach initially utilizes a combination of the method of weighted residuals with empirical eigenfunctions as basis functions to discretize the spatial domain and derive ODE models that accurately describe the dominant process dynamic behavior. Subsequently, control vector parametrization is used to discretize in time the resulting infinite-dimensional optimization problem and obtain a finite dimensional algebraic nonlinear program, which can be solved utilizing standard nonlinear programming techniques. We use the proposed approach to compute optimal process operating profiles to achieve radially uniform GaN thin-films in a vertical MOVPE reactor with showerhead configuration, based on the observation that a change in precursor distribution across the reactor inlet results in an altered GaN deposition rate profile over the substrate. We demonstrate, through simulations, that by *optimally* changing the precursors' concentration across the reactor inlet, it is possible to reduce the radial thickness nonuniformity of the deposited GaN film from a nominal 33% (for a time-constant operation) to 3.1% (under the proposed time-varying operation).

Problem Formulation

We focus on spatially-distributed processes modeled by highly dissipative PDE systems with the following state-space description:

$$\frac{\partial x}{\partial t} = \mathscr{A}(x) + f(t, x, d),\ x(z, 0) = x_0(z)$$

$$g\left(x, \frac{dx}{d\eta}, \ldots, \frac{d^{n_o-1}x}{d\eta^{n_o-1}}\right) = 0, \quad \text{on } \Gamma \tag{1}$$

where $x(z, t) \in \mathbb{R}^n$ denotes the vector of state variables, $t \in [0, t_f]$ is the time (t_f is the terminal time), $z = [z_1, z_2, z_3] \in \Omega \subset \mathbb{R}^3$ is the vector of spatial coordinates, Ω is the domain of definition of the process, and Γ its boundary. $\mathscr{A}(x)$ is a dissipative, possibly nonlinear, spatial differential operator which includes higher-order spatial derivatives, $f(t, x, d)$ is a nonlinear, possibly time-varying, vector function which is assumed to be sufficiently smooth with respect to its arguments, $d(t) \in \mathbb{R}^p$ is the vector of design variables which are assumed to be piecewise continuous functions of time, $h(x, dx/d\eta, \ldots, d^{n_o-1}x/d\eta^{n_o-1})$ is a nonlinear vector function, which is assumed to be sufficiently smooth (n_o, an even number, is the order of the PDE of Eq. 1), $dx/d\eta|_\Gamma$ denotes the derivative in the direction perpendicular to the boundary and $x_0(z)$ is a smooth vector function of z.

The system of Eq. 1 arises in the modeling of a wide range of dynamic spatially distributed processes including both transport-reaction processes and several classes of dissipative fluid dynamic systems.[10] The nonlinear structure of the spatial differential operator, $\mathscr{A}(x)$, allows accounting for the explicit dependence of diffusivity and thermal conductivity on temperature and concentration in certain transport-reaction processes, while the nonlinear term $f(t, x, d)$ allows modeling complex reaction mechanisms, as we will illustrate in the application section for the GaN thin film epitaxy process.

A general optimization problem for the system of Eq. 1 can be formulated as follows

$$\min \int_0^{t_f} \int_\Omega G(x(z, t), d(t)) dz dt$$

$$s.t.$$

$$-\frac{\partial x}{\partial t} + \mathscr{A}(x) + f(t, x, d) = 0,$$

$$x(z, 0) = x_0(z),\ g\left(x, \frac{dx}{d\eta}, \ldots, \frac{d^{n_o-1}x}{d\eta^{n_o-1}}\right) = 0 \quad \text{on } \Gamma$$

$$g(x, d) \leq 0,\ \forall\, z \in \Omega,\ t \in [0, t_f] \tag{2}$$

where $\int_0^{t_f} \int_\Omega G(x, d) dz dt$ is the objective functional, and $g(x, d)$ is the vector of inequality constraints which may include bounds on the state and design variables. Both $G(x, d)$ and $g(x, d)$ are assumed to be sufficiently smooth functions of their arguments.

Spatial discretization

It is desired to obtain a finite dimensional approximation of the infinite dimensional program developed above through spatial discretization of the imposed PDE constraints. The reader may refer to [30,31] for alternative methods for semi-infinite programming. To formulate low dimensional NLPs, we employ the method of weighted residuals with empirical eigenfunctions as basis functions[6,4,5] instead of standard finite difference or finite elements. The rationale behind this approach is that solutions of highly dissipative PDEs are dominated by a finite (typically small) number of degrees of freedom.[11] For the case of parabolic PDEs with linear differential operators, these can be identified as finite dimensional slow eigenmodes,[10] and can be calculated analytically, however for nonlinear differential operators with spatially varying coefficients, such an analytical solution is, in general, not feasible. Karhunen-Loève expansion, coupled with the method of snapshots, is an attractive alternative in such situations. In the following, we briefly review the method of weighted residuals which is followed by a brief description of Karhunen-Loève expansion.

Method of weighted residuals

We derive finite-dimensional approximations of the infinite-dimensional nonlinear program of Eq. 2 by using the method of weighted residuals. To simplify the notation, we consider the optimization program of Eq. 2 with $n = 1$. In principle, $x(z, t)$ can be represented as an infinite series in terms of a complete set of basis functions $\phi_k(z)$. We can obtain an approximation $x_N(z, t)$, by truncating the series expansion of $x(z, t)$ up to order N, as follows

$$x_N(z, t) = \sum_{k=1}^{N} a_{kN}(t)\phi_k(z) \xrightarrow{N \to \infty} x(z, t) = \sum_{k=1}^{\infty} a_k(t)\phi_k(z) \tag{3}$$

where $a_{kN}(t)$, $a_k(t)$ are time-varying coefficients.

Substituting the expansion of Eq. 3 into Eq. 2, multiplying the PDE and the inequality constraints with the weighting functions $\psi_\nu(z)$, and integrating over the entire spatial domain, the following finite-dimensional dynamic nonlinear program with ODE equality constraints is obtained, where the optimization parameters are the design variables $d(t)$, and the time varying coefficients $a_{kN}(t)$

$$\min \int_0^{t_f} \int_\Omega G\left(\sum_{k=1}^{N} a_{kN}(t)\phi_k(z), d\right) dz dt$$

$$s.t.$$

$$-\sum_{k=1}^{N} \dot{a}_{kN}\left(\int_\Omega \psi_\nu(z)\phi_k(z)dz\right) + \int_\Omega \psi_\nu(z)\mathscr{A}\left(\sum_{k=1}^{N} a_{kN}(t)\phi_k(z)\right)dz$$

$$+ \int_\Omega \psi_\nu(z) f\left(t, \sum_{k=1}^{N} a_{kN}(t)\phi_k(z), d\right)dz = 0$$

$$\int_\Omega \psi_\nu(z) g\left(\sum_{k=1}^{N} a_{kN}\phi_k(z), d\right)dz \leq 0 \tag{4}$$

where $a_{kN}(t)$ is the approximation of $a_k(t)$ obtained by an N-th order truncation. From Eq. 4, it is clear that the form of the algebraic equalities and inequalities depend on the choice of the weighting functions, as well as on N. Owing to the smoothness of the functions $G(x, d)$, $\mathscr{A}(x)$, $f(t, x, d)$, $g(x, d)$, and the completeness of the set of basis functions $\phi_k(z)$, the nonlinear program of Eq. 4 is a well-defined approximation of the infinite-dimensional program of Eq. 2 in the sense that the optimal solution of the program of Eq. 4 converges to the optimal solution of the program of Eq. 2 as $N \rightarrow \infty$.

Computation of empirical eigenfunctions via Karhunen-Loève expansion

In this section, we use the solution data of the system of Eq. 1 to construct global basis functions using Karhunen-Loève (KL) expansion. The motivation for studying this approach is provided by the occurrence of dominant spatial patterns in the solution of several dissipative PDEs, which should be accounted for in the shape of the global basis functions. This approach will be useful in the context of systems of dissipative PDEs that involve nonlinear spatial differential operators and spatially-varying coefficients that lead to nonsymmetric solution profiles. KL expansion is a procedure used to compute an optimal set of empirical eigenfunctions from an appropriately constructed set of solutions of the PDE system of Eq. 1, obtained from high-order discretizations (for example, using standard packages or process data directly). In this work, the ensemble of solutions is constructed by computing the solutions of the PDE system of Eq. 1 for different values of $d(t)$, and different initial conditions. Specifically, we construct a representative ensemble using the following procedure (see also [32,5] for a detailed discussion on ensemble construction):

- First, we create a set of different initial conditions.
- We then discretize the interval in which each design variable d_m ($m = 1, \ldots, p$) is constrained to be into m_{d_m} (not necessarily equispaced) subintervals. The discrete values of d_m are denoted by $d_{m,j}$, $j = 1, \ldots, m_{d_m} - 1$.
- We also discretize the time-interval into n_{d_m} time subintervals (also not necessarily equispaced).
- Subsequently, we compute a set of time profiles for each of the design variables $d_m(t)$ by assigning values for $d_m(t)$ at different time instants t_j, $d_{m,j}$, and subsequently computing $d_m(t)$ for the entire time interval of process operation using linear interpolation.
- Finally, we compute an ensemble of PDE solution data for all possible combinations of initial conditions and profiles of $d(t)$.

Application of KL expansion to this ensemble of data provides an orthogonal set of basis functions (known as empirical eigenfunctions) for the representation of the ensemble, as well as a measure of the relative contribution of each basis function to the total energy (mean square fluctuation) of the ensemble. A truncated series representation of the ensemble data in terms of the dominant basis functions has a smaller mean square error than a representation by any other basis of the same dimension.[33] This implies that the projection on the subspace spanned by the empirical eigenfunctions will on average contain the most energy possible compared to all other linear decompositions, for a given number of modes. Therefore, the KL expansion yields the most efficient way for computing the basis functions (corresponding to the largest empirical eigenvalues) capturing the dominant patterns of the ensemble.

For simplicity of the presentation, we describe the KL expansion in the context of the system of Eq. 1 with $n = 1$, and assume that there is available a sufficiently large set of solutions of this system for different values of d, $\{\bar{v}_\kappa\}$, consisting of K sampled states, $\bar{v}_\kappa(z)$, (which are typically called “snapshots”). The following analysis is largely adopted from [6] and is reproduced here for completeness. The reader may refer to [34,33,8,9] for a detailed presentation and analysis of the KL expansion. We define the ensemble average of snapshots as $\langle \bar{v}_\kappa \rangle := 1/K \sum_{\kappa=1}^{K} \bar{v}_\kappa(z)$ (we note that non-uniform sampling of the snapshots and weighted ensemble average can be also considered; see, for example, [32]). Furthermore, the ensemble average of snapshots $\langle \bar{v}_\kappa \rangle$ is subtracted out from the snapshots that is,

$$v_\kappa = \bar{v}_\kappa - \langle \bar{v}_\kappa \rangle \tag{5}$$

so that only fluctuations are analyzed. It is useful to analyze these fluctuations rather than the actual variables because usually fewer eigenfunctions are required to fit them accurately.[8] The issue is how to obtain the most typical or characteristic structure (in a sense that will become clear below) $\phi(z)$ among these snapshots $\{v_\kappa\}$. Mathematically, this problem can be posed as the one of obtaining a function $\phi(z)$ that maximizes the following objective function

$$\text{Maximize} \frac{\langle (\phi, v_\kappa)^2 \rangle}{(\phi, \phi)}$$

$$s.t.\ (\phi, \phi) = 1,\ \phi \in L^2([\Omega]) \tag{6}$$

which, other words, implies that the projection of $\bar{v}_k$ on the subspace spanned by $\phi(z)$ captures maximum energy. Here, (x, y) denotes complex inner-product defined as

$$(x, y) = \int_\Omega \bar{x}(z) y(z) dz \tag{7}$$

The constraint $(\phi, \phi) = 1$ is imposed to ensure that the function, $\phi(z)$, computed as a solution of the above maximization problem, is unique. An alternative way to express the constrained optimization problem of Eq. 6 is to solve for ϕ such that

$$\frac{d\bar{L}(\phi + \delta\psi)}{d\delta}(\delta = 0) = 0,\ (\phi, \phi) = 1 \tag{8}$$

where $\bar{L} = \langle (\phi, v_\kappa)^2 \rangle - \lambda((\phi, \phi) - 1)$ is the corresponding Lagrangian functional, and δ is a real number.

Using the definitions of inner product and ensemble average, $d\bar{L}(\phi + \delta\psi)/d\delta(\delta = 0)$ can be computed from the following expression

$$\frac{d\bar{L}(\phi + \delta\psi)}{d\delta}(\delta = 0) = \int_\Omega \left(\left\{ \int_\Omega \langle v_\kappa(z)v_\kappa(\bar{z})\rangle \phi(z)dz \right\} - \lambda\phi(\bar{z}) \right) \psi(\bar{z})d\bar{z} \quad (9)$$

Since $\psi(\bar{z})$ is an arbitrary function, the necessary conditions for optimality take the form

$$\int_\Omega \langle v_\kappa(z)v_\kappa(\bar{z})\rangle \phi(z)dz = \lambda\phi(\bar{z}),\ (\phi,\ \phi) = 1 \quad (10)$$

Introducing the two-point correlation function

$$K(z,\ \bar{z}) = \langle v_\kappa(z)v_\kappa(\bar{z})\rangle = \frac{1}{K}\sum_{\kappa=1}^{K} v_\kappa(z)v_\kappa(\bar{z}) \quad (11)$$

and the linear operator

$$R := \int_\Omega K(z, \bar{z})d\bar{z} \quad (12)$$

the optimality condition of Eq. 10 reduces to the following eigenvalue-eigenfunction problem of the integral operator

$$R\phi = \lambda\phi \Rightarrow \int_\Omega K(z,\ \bar{z})\phi(\bar{z})d\bar{z} = \lambda\phi(z) \quad (13)$$

The computation of the solution of the above integral eigenvalue problem is, in general, a very expensive computational task. To circumvent this problem, Sirovich, in 1987, introduced the method of snapshots.[8,9] The central idea of this technique is to assume that the requisite eigenfunction, $\phi(z)$, can be expressed as a linear combination of the snapshots that is,

$$\phi(z) = \sum_k c_k v_k(z) \quad (14)$$

Substituting the above expression for $\phi(z)$ on Eq. 13, we obtain the following eigenvalue problem

$$\int_\Omega \frac{1}{K}\sum_{\kappa=1}^{K} v_\kappa(z)v_\kappa(\bar{z}) \sum_{k=1}^{K} c_k v_k(\bar{z})d\bar{z} = \lambda \sum_{k=1}^{K} c_k v_k(z) \quad (15)$$

Defining

$$B^{\kappa k} := \frac{1}{K}\int_\Omega v_\kappa(\bar{z})v_k(\bar{z})d\bar{z} \quad (16)$$

the eigenvalue problem of Eq. 15 can be equivalently written as

$$Bc = \lambda c \quad (17)$$

The solution of the above eigenvalue problem (which can be obtained by utilizing standard methods from linear algebra) yields the eigenvectors $c = [c_1 \cdots c_K]$ which can be used in Eq. 14 to construct the eigenfunction $\phi(z)$. From the structure of the matrix B, it follows that it is symmetric and positive semi-definite, and, thus, its eigenvalues λ_κ, $\kappa = 1, \ldots, K$, are real and non-negative. The relative magnitude of the eigenvalues represents a measure of the fraction of the "energy" embedded in the ensemble captured by the corresponding eigenfunctions. Furthermore, the resulting eigenfunctions form an orthogonal set, that is,

$$\int_\Omega \phi_i(z)\phi_j(z)dz = 0,\ i \neq j \quad (18)$$

Remark 1: The value of m_{d_m} should be determined based on the effect of the design variable d_m on the solution of the system of Eq. 1 (if, for example, the effect of the variable d_1 is larger that the effect of the variable d_2, then m_{d_1} should be larger than m_{d_2}).

Remark 2: It should be noted that the kernel in Eq. 10 is not symmetric for cylindrical or spherical geometries.[21] However, the reformulated problem given by Eq. 17 is symmetric irrespective of spatial geometry.

Remark 3: The basis that we compute using KL expansion is specific to the process under investigation and independent of the specific optimization problem we try to solve. Therefore, the same basis can be used to perform computationally efficient optimizations with respect to different objective functionals associated with the same underlying set of partial differential equations.

Remark 4: Even though it is expected that the use of more basis functions in the series expansion of Eq. 3 would improve the accuracy of the computed approximate model of Eq. 4, the use of empirical eigenfunctions corresponding to very small eigenvalues should be avoided because such eigenfunctions are contaminated with significant round-off errors.

Remark 5: Iterative methods, such as Krylov subspace methods can be used to reduce the computational cost associated with the computation of the system eigenvalues and eigenfunctions.

Temporal Discretization

The computation of the solution of infinite optimization problems usually involves a reformulation step discretizing the infinite variable domain (with the exception of approaches based on calculus of variations). In the current section, we discretize the infinite temporal domain of the dynamic nonlinear program to obtain a finite dimensional NLP which can be subjected to subsequent numerical solution.

Discretization in the temporal domain can be carried out for both state and design variables using finite differences, orthogonal collocation and so on, to obtain discrete versions of equality and inequality constraints and cost functional. Optimization, then, is carried out in the full space of discretized variables which often require large computational resources and wall-clock time. In order to reduce the dimensionality of the NLP, control vector parametrization (CVP) can be employed, where only the design variables are discretized. Equality constraints are directly integrated in time, within the bounds that are set by the inequality constraints. CVP, hence, requires the optimization to be performed only in the space of design variables which often reduces the computational load.

The most straightforward way is to partition the temporal domain into m_t intervals and express $d(t)$ as piecewise constant function over the entire time domain. Hence, the vector function $d(t)$ is expressed as a series of the form

$$d(t) = \sum_{i=0}^{m_t-1} d_{i+1}[H(t - t_i) - H(t_{i+1} - t)] \tag{19}$$

where $H(\cdot)$ is the standard Heaviside function. This would result in a set of ODE equality and inequality constraints for m_t time intervals $\delta t_i = t_i - t_{i-1}$, $\forall i = 1, \ldots, m_t$. CVP requires direct integration of these constraints, either analytically (if possible) or numerically, and d_i and δt_i are the optimization variables. Application of CVP to the dynamic nonlinear program of Eq. 4 results in an algebraic nonlinear program of dimension $m_t \times (p + 1)$, which has the following general form

$$\begin{aligned} &\min F(x) \\ &s.t. \\ &h(x) = 0 \\ &g(x) \leq 0 \end{aligned} \tag{20}$$

where the explicit form of the functions $F(x)$, $h(x)$, $g(x)$ is omitted for brevity. Note that in the earlier formulation the integrated ODEs reside in $h(x)$. Note that discretization using finite differences (backward or forward) for all variables would have produced a NLP of dimension $m_t \times (p + N)$.

Remark 6: Referring to the parametrization of the control vector, even though we have expressed it as piecewise constant function, other representations, such as piecewise polynomial[13] or Lagrange polynomials[15] have also been used. Such schemes can be employed with the proposed spatial discretization in a straightforward manner.

Remark 7: It might be required to enforce some continuity between d_i and d_{i+1} in order to account for actual controller dynamics, when they are comparable to process dynamics. Refer to [15] for a detailed discussion.

Computation of optimal solution

In this section, we propose a computationally efficient procedure for the computation of an accurate optimal solution of the infinite dimensional nonlinear program of Eq. 1 using standard optimal search algorithms, such as successive quadratic programming (SQP), Broyden, Fletcher, Goldfarb, Shanno (BFGS), and Luus-Jakkola (LJ) algorithms.[35,36] The validity of the optimal solution computed is investigated by checking convergence to a specific optimum as N and m_t increase.

We formulate the procedure used for the computation of the optimal solution of the infinite-dimensional program of Eq. 2 in following algorithm:

- Step 1: Compute an initial guess for N, $\hat{N}$, based on the magnitude of the eigenvalues corresponding to the eigenfunctions.
- Step 2: Use the spatial and temporal discretization procedures of sections entitled “Problem Formulation” and “Spatial Discretization,” respectively, to derive a finite-dimensional program of the form of Eq. 20.
- Step 3: Solve the resulting finite-dimensional program using standard search algorithms to compute an optimal solution.
- Step 4: Derive and solve a new finite-dimensional program of the form of Eq. 20 by performing spatial discretization with $N = \hat{N} + 1$.
- Step 5: Compare the two optimal solutions for $N = \hat{N}$ and $N = \hat{N} + 1$. If they are close (according to the desired accuracy metric), then stop; a convergent optimal solution has been found. If not, then go back to step 2, and perform spatial discretization with $N = \hat{N} + 2$.
- Step 6: Reduce the temporal discretization step δt to increase the resolution in the temporal domain.

The structure of the earlier algorithm is motivated by the fact that the discrepancy between the infinite-dimensional program, and its finite-dimensional approximation of Eq. 4 decreases as the number of basis functions N, used in the expansion of Eq. 3 increases (at least, up to the point where round-off errors become significant). This is a consequence of the hierarchy of the eigenfunctions. Also, the increase in the computational cost for the solution of the optimization problem due to the above iterative scheme can be minimized by judiciously choosing the initial guess for N, and employing the optimal solution obtained at *i-th iteration as initial guess for the* $i + 1$*-th* iteration. In the current case, the initial number of eigenfunctions was chosen such that they accounted for at least 99% of the energy of the ensemble of snapshots.

In the next section, we implement the presented reduced-order optimization approach to the thin film epitaxy of *GaN*. Our focus is to obtain a high degree of spatial uniformity in the thin films that are grown through metallorganic vapor phase epitaxy (MOVPE).

Application to GaN Thin Film Epitaxy

Process description and modeling

Currently GaN on Sapphire, Si or SiC substrates is produced in a two stage MOVPE process usually with trimethylgallium (TMGa) and NH_3 as precursors for gallium and nitrogen respectively, diluted in H_2 (carrier gas).[37,38] During the first stage, a GaN nucleation layer is formed on the wafer surface at low-temperatures (600°C), forming a buffer layer between the GaN epilayer and the Sapphire substrate. The need for the buffer layer originates from the large lattice mismatch between Sapphire and GaN (13%). Alternatively, AlN is also employed as a buffer layer.[38] At the termination of the first stage the reactor is purged with carrier gas and the substrate temperature is increased, annealing the nucleation layer. The rate of temperature increase has been found to be significant, with a rate

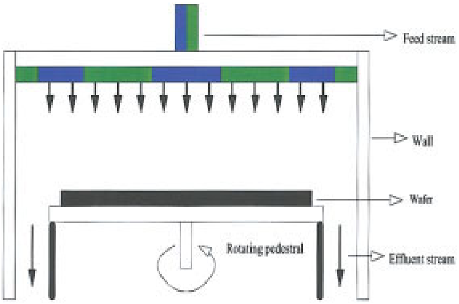

Figure 1. Vertical MOVPE reactor with a three concentric ring showerhead configuration.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

of 40°C/min to be optimal in terms of the resulting GaN layer quality.[39] The second stage of the GaN epitaxy is initiated at temperature of 1060°C, forming the desired GaN epilayer.

The growth rate and the structural properties of the thin film are controlled by multiple fundamental phenomena that occur during the process cycle, including gas-phase reactions and transport of the precursor gases within the reactor and adsorption, surface diffusion, surface-reactions and desorption of the adsorbed species.[37] Parasitic prereactions between *Ga* and *N* precursors that form Lewis acid-base adducts are known to occur, which, on one hand, deplete the feedstream of limiting species and, on the other hand, can negatively affect the film quality owing to deposition of adduct on cold reactor walls, which leads to particulate formation.[40] To avoid this problem, a vertical showerhead reactor with multiple inlets is used, and gallium and nitrogen containing precursors are fed from different inlets, so that mixing between the two occurs just above the wafer.[41]

A vertical MOVPE reactor with showerhead configuration is shown in Figure 1. Precursor gases for gallium and nitrogen enter through distinct inlets directly above the heated substrate over which the film is deposited. In order to improve the radial mixing of gas-streams, the substrate may be rotated with the help of a pedestal, situated below the substrate. $TMGa$ and NH_3 diluted in hydrogen carrier gas were used as precursors for gallium and nitrogen, respectively. The split inlet design comprising of three concentric rings (which is primarily used to avoid pre-reactions between precursors) allows for the spatial variation in concentration of the precursors across the inlet. Further details of process conditions and reactor geometry are provided in Table 1.

Table 1. Process Conditions and Reactor Geometry

Reactor radius	2 in
Substrate radius (R_s)	1.5 in
Number of inlets	3
inner inlet outer radius	0.5 in
middle inlet outer radius	1 in
outer inlet outer radius	1.5 in
Substrate to inlet distance (z_0)	3 in
Reactor pressure	0.1 atm
Reactor wall temperature	300 K
Substrate temperature (T_s)	1300 K
Inlet temperature	300 K
Inlet velocity	80 cm/s
X^*_{TMGa}	1.5×10^{-4}
$X^*_{\mathrm{NH_3}}$	0.15

*Inlet mole fractions of reactant in H_2 carrier gas.

Table 2. Gas Phase Reactions

Reaction	k_0	E
(G1) $Ga(CH_3)_3 \rightarrow Ga(CH_3)_2 + CH_3$	3.5×10^{15}	59.5
(G2) $Ga(CH_3)_2 \rightarrow GaCH_3 + CH_3$	8.7×10^{7}	35.4
(G3) $Ga(CH_3)_3 + NH_3 \rightarrow (CH_3)_3Ga : NH_3$	—*	0
(G4) $(CH_3)_3Ga : NH_3 \rightarrow Ga(CH_3)_3 + NH_3$	1×10^{14}	18.5
(G5) $(CH_3)_3Ga : NH_3 \rightarrow (CH_3)_2Ga : NH_2 + CH_4$	1×10^{14}	49

*Rate determined from bimolecular collision rate.

The reaction model describing the reactions between gas-phase species and gas-surface reactions has been adopted from [29,42] and is shown in Tables 2 and 3. Reactions $G1$ and $G2$ describe the gas-phase decomposition of *TMGa,* and dimethyl gallium (DMGa) respectively. Reaction $G3$ describes the recombination reaction between *TMGa,* and ammonia to form an adduct whose rate is estimated by the rate of bimolecular collisions,[29] which according to kinetic theory is $k = \pi\sigma_{AB}^2(8k_BT/\pi\mu)^{0.5}$, where k_B is the Boltzmann's constant, T is the absolute temperature in gas phase, and μ is the reduced mass given by $1/\mu = 1/m_{TMGa} + 1/m_{NH_3}$, where m_{TMGa} and m_{NH_3}, are the molecular weights of *TMGa,* and ammonia, respectively. The mean collision diameter (σ_{AB}) for two species is given by $\sigma_{AB} = 1/2 \times (\sigma_A + \sigma_B)$, where σ_A and σ_B are the collision diameters of A and B. $G4$ and $G5$ are adduct dissociation and methane elimination reactions, respectively. Some investigations have taken into account the formation of cyclic trimer from $(CH_3)_2Ga : NH_2$ into the gas phase chemistry,[43,44] but recently, the trimer's concentration has been shown to be negligible and, thus, is neglected from the gas-phase kinetic model.[45] The rates of adsorption reactions $S1$–$S3$ were calculated assuming the sticking coefficient to be equal to unity. The rate of nitrogen adsorption $S4$ on the surface was set to be equal to the combined rate of $S1$, $S2$ and $S3$ in order to maintain film stoichiometry. Reactions $S5$ and $S6$ are included for completeness, though simulations revealed that their contribution towards the overall film growth is negligible.

The earlier reaction scheme was incorporated into momentum, energy and mass conservation equations to provide with a process model consisting of $N_s + 3$ PDEs of the following form

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0$$

$$\frac{\partial (\rho \mathbf{u})}{\partial t} + \nabla \cdot (\rho \mathbf{u}\mathbf{u}) - \nabla \cdot \mathbf{T} - \rho \mathbf{g} = 0$$

Table 3. Gas-Surface Reactions

Reaction†	s‡	E
(S1) $Ga(CH_3)_3 + S \rightarrow Ga(bulk) + 3CH_3$	1	0
(S2) $Ga(CH_3)_2 + S \rightarrow Ga(bulk) + 2CH_3$	1	0
(S3) $GaCH_3 + S \rightarrow Ga(bulk) + CH_3$	1	0
(S4) $NH_3 + S \rightarrow N(bulk) + CH_3$	—*	—*
(S5) $(CH_3)_3Ga : NH_3 + 2S \rightarrow GaN + 3CH_4$	1	0
(S6) $(CH_3)_2Ga : NH_2 + 2S \rightarrow GaN + 2CH_4$	1	0

*Rate equal to S1 + S2 + S3.
†S denotes a free surface site.
‡$s = 1$ denotes a unity sticking coefficient at zero coverage.

$$C_p\left[\frac{\partial(\rho T)}{\partial t} + \nabla\cdot(\rho\mathbf{u}T)\right] = -\nabla\cdot\mathbf{q} - \sum_k h_k W_k \dot{\omega}$$

$$\frac{\partial(\rho Y_k)}{\partial t} + \nabla\cdot(\rho\mathbf{u}Y_k) = -\nabla\cdot\mathbf{j}_k + W_k\dot{\omega}_k,\ \forall\, k = 1,\ldots,N_s - 1 \tag{21}$$

where ρ is the density of the mixture, $\mathbf{u}$ is the fluid velocity vector, $\mathbf{T}$ is the stress tensor, C_p is the specific heat capacity of the multicomponent mixture, T is the temperature, $\mathbf{q}$ is the heat flux due to conduction and h_k, W_k and Y_k are the partial specific enthalpy, molecular weight and mass fractions of species involved. $\dot{\omega}_k$ and $\mathbf{j}_k$ are the net production rate due to homogeneous reactions and mass flux respectively of species k out of the total N_s gaseous species. Physical properties of the gas phase mixture such as viscosity, thermal conductivity, binary diffusion coefficients and specific heat capacities were calculated from kinetic theory and were a function of composition and temperature. Since NH_3 and H_2 were present in considerable mass fractions, the mass balance equations were no longer independent and full multicomponent diffusion model was used for species diffusion.

The earlier mathematical model was solved for cylindrical coordinates assuming axisymmetric conditions (no azimuthal dependence of any process variable). Simulations were performed using the commercially available software FLUENT 6.02. In particular, we were interested in variations in the process variable profiles that might arise by varying the precursor flowing through the reactor inlets. In the following subsection we present these results.

Effect of inlet configuration

As hydrogen is the carrier gas, the possible gas streams that can flow through any of the reactor inlets are: $TMGa + H_2$ (which we will denote as *T*), $NH_3 + H_2$ (denoted as N) and pure H_2 (denoted as H). Based on the above available choices of gas streams, 5 distinct *reactor inlet configurations* (characterized by the choice of precursor flowing through the 3 inlets) were investigated, namely *TNN, NTH, TNT, TNH* and *THN.*[1] In Figure 2, concentration profiles of important gas phase species inside the reactor are presented for the *TNN* inlet configuration. It can be seen that only *DMGa* and *MMGa* are present in significant mass fractions over the wafer surface, while *TMGa* and adduct are consumed away from the wafer surface, and, hence, they are the only species responsible for GaN film growth. Similar observations are made for the rest of the inlet configurations, however, the concentration of these species across the reactor is dependent on the inlet configuration (for brevity, we do not present the concentration profiles for the rest of the inlet configurations). This results in a characteristic *Ga* deposition rate profile across the wafer for each inlet configuration (shown in Figure 3). This motivated us to search for an optimal switching criteria for inlet configurations which would result in a final film which has *minimal* radial thickness nonuniformity. It is important to note here that the ratio of dilution of *Ga* and *N* precursors in the carrier gas has not been altered for any of the inlet configurations, as the quality of the *GaN* film is experimentally found to be sensitive to it.

Simulation studies allow the following simplifying assumptions in the formulation of the optimization problem. In Figures 4 and 5 we present mass fractions of CH_4 and $(CH_3)_2Ga : NH_2$ across the reactor for the *TNN* inlet configuration. It can be seen that these species are present in negligible proportion as compared to other gas phase species (compare with mass fractions of *TMGa, DMGa, MMGa,* and so on). The same observation was made for the rest of inlet configurations.

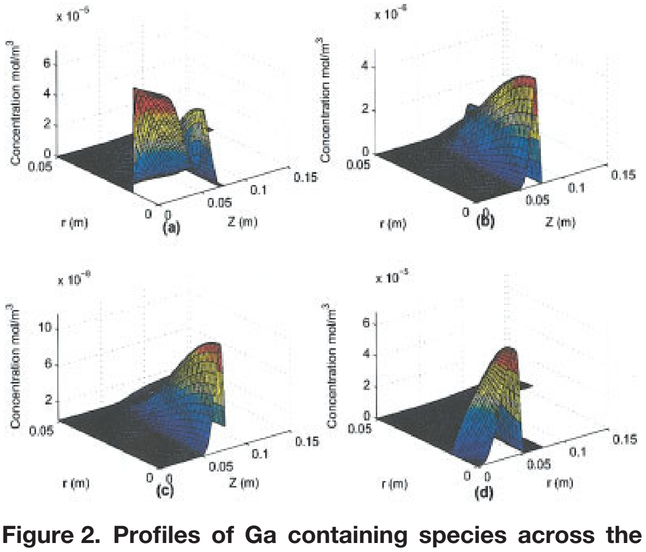

Figure 2. Profiles of Ga containing species across the reactor for *TNN* inlet configuration. (a) *TMGa,* (b) *DMGa,* (c) *MMGa,* and (d) *adduct.*

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

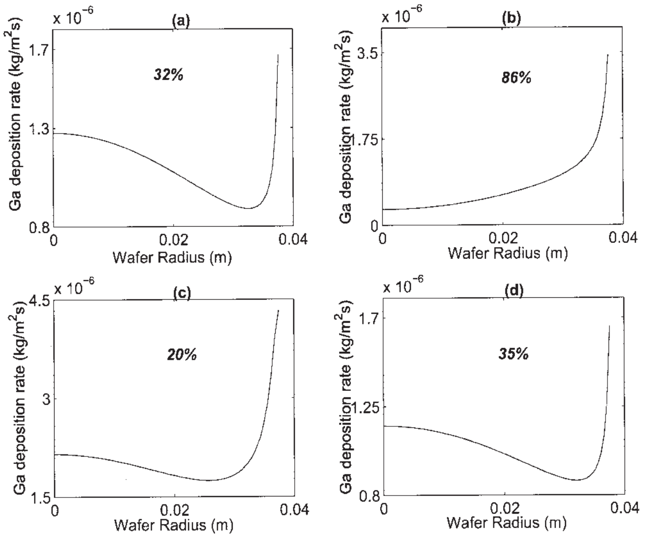

Figure 3. Steady-state gallium deposition-rate profile across the substrate surface for inlet configurations (a) *TNN,* (b) *NTH,* (c) *TNT,* (d) *TNH,* respectively.

Corresponding thickness nonuniformities are also shown for cutoff radius = 0.03429 m.

[1] For more clarity, *TNN* for example, implies $TMGa + H_2$ flowing through the innermost inlet and $NH_3 + H_2$ flowing through the middle and outermost inlets.

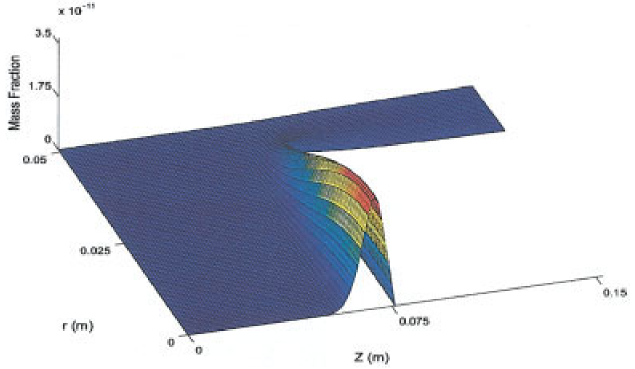

Figure 4. Mass fraction of $(CH_3)_2Ga$: NH_2 for *TNN* configuration.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Hence, these species and the corresponding gas-phase and surface reactions (*G*5, *S*5 and *S*6) were omitted. Similarly, it was found that the variations in axial and radial velocities were small irrespective of the inlet configuration. Thus, time-invariant axial and radial velocities, equal to their ensemble average, were employed during the formulation of reduced-order process model. Also, heat generation, due to chemical reactions, was ignored because of low concentration of reacting species (for example, maximum *TMGa* mole fraction was 1.5×10^{-4}).

An analysis of the thermal Peclet number revealed that convective heat transfer was small compared to heat transfer by conduction (characterized by low values of Peclet number), which allowed us to neglect the convective heat-transfer term in the energy conservation equation. Also the dependence of specific heat and thermal conductivity of the mixture on temperature and mixture composition was found to be analogous. These considerations allowed the energy equation to be decoupled from the rest of the conservation equations. KL expansion identified a single dominant eigenfunction, whose eigenvalue captured more than 98% of the energy included in the ensemble of snapshots, which allowed us to assume an exponentially decaying relationship for the deviation of the temperature from the steady-state profile. The time constant was estimated from physical considerations. The simulations also revealed that the deviation of the temperature spatial profile from the steady-state tends to die out quickly after switching from one inlet configuration to another, which confirmed the earlier argument.

Reduced-order modeling and optimization

As mentioned earlier, different inlet configurations can be employed to obtain different distributions of precursors across the reactor inlet. Switching from one inlet configuration to another causes the system to dynamically evolve to a new steady-state with a characteristic deposition rate profile. Under the objective of minimum nonuniformity in the final film thickness, the goal of optimization is to ascertain an optimal switching policy for inlet configurations.

The equality constraints for optimization are provided by mass, momentum and energy conservation laws and assume the form of coupled PDEs in the present case. A finite dimensional approximation of this infinite-dimensional problem can be obtained through discretization in space and time. We avoid the use of standard finite differences for discretization due to the large size of the resulting NLP. Instead, we adopted the approach described in the section entitled "Problem Formulation," and performed spatial discretization using the method of weighted residuals using empirical eigenfunctions (obtained by Karhunen-Loève expansion) as basis functions. Temporal discretization was performed using control vector parametrization. Simulation data from Fluent for a variety of inlet configurations (initial conditions) were employed as "snapshots" to construct empirical eigenfunctions that describe the dominant spatial patterns in the solution of the PDEs for momentum, energy and mass transport. To capture the system dynamics, 28 snapshots were obtained after each switching, and a total of 25 different switchings (from six different inlet configurations) were employed to generate an ensemble of 28 × 25 snapshots.

Using these snapshots, we computed empirical eigenfunctions for the concentrations of five species namely, *TMGa, DMGa, MMGa, adduct* and NH_3 to derive a reduced-order model. From the ensemble of snapshots a distinct set of 16 eigenfunctions (which accounted for at least 99% of the energy embedded in the ensemble of snapshots) was computed for each switching to reproduce the spatiotemporal behavior of the concentration of the species listed above. In effect, this led to a separate ODE model for each switching of inlet configuration. Thus, the derived reduced-order model based on the computed empirical eigenfunctions, involved 64 ODEs (for four switchings employed during optimization) to describe the dynamic behavior of the concentration of these species with axial and radial velocities held constant at their respective ensemble averages and an algebraic expression describing the (exponentially decaying) evolution of the spatial variations of the temperature from its steady state. The leading eigenfunctions for *TMG, DMG* and *MMG* for different inlet configurations are shown in Figures 6, 7 and 8, respectively.

The optimization problem was formulated as follows

$$\min F = w_1 \int_0^{R_o} \left\{ \int_0^{t_f} (R_{dep}(r, t)dt - \bar{H}(\delta t))dt \right\}^2 dr + w_2[H_{obj} - \bar{H}(\delta t)]^2$$

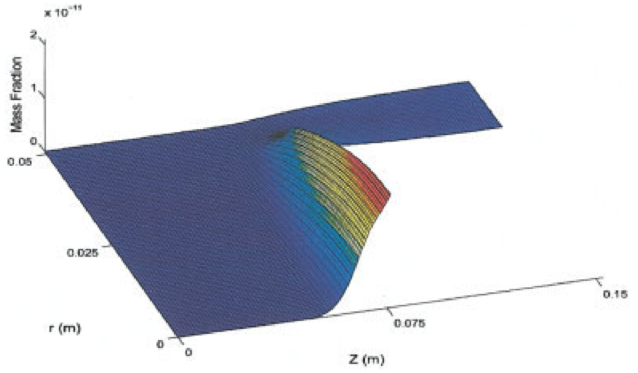

Figure 5. Mass fraction of CH_4 for *TNN* configuration.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

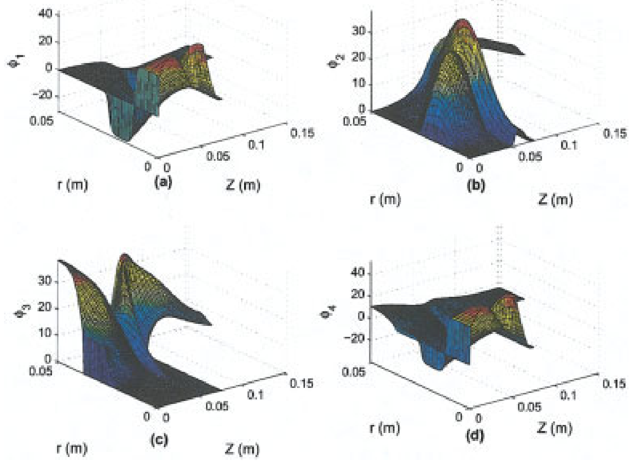

Figure 6. Dominant eigenfunctions for concentration of *TMGa* for inlet configurations (a) *TNN,* (b) *NTH,* (c) *TNT,* and (d) *TNH,* respectively.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$$s.t.$$

$$\bar{H} = \frac{1}{R_o}\int_0^{R_o}\int_0^{t_f} R_{dep}(r, t)dtdr$$

$$R_{dep}(r, t) = \sum_{ls} k_{ls}(T_s)C_{ls}(t, r, z = z_0)$$

$$\mathbf{u} = u^r(r, z, t)\mathbf{e}_r + u^z(r, z, t)\mathbf{e}_z = \mathbf{e}_r \sum_{i=1}^{n_{u^r}} a_i^{u^r}(t)\psi_i^{u^r}(r, z) + \mathbf{e}_z \sum_{i=1}^{n_{u^z}} a_i^{u^z}(t)\psi_i^{u^z}(r, z)$$

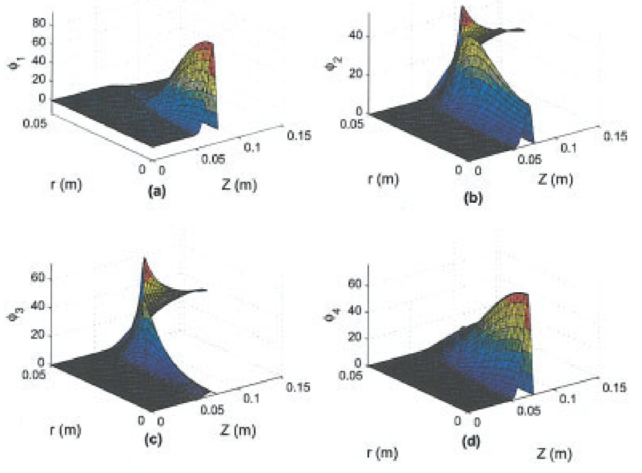

Figure 7. Dominant eigenfunctions for concentration of *DMGa* for inlet configurations (a) *TNN,* (b) *NTH,* (c) *TNT,* (d) *TNH,* respectively.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

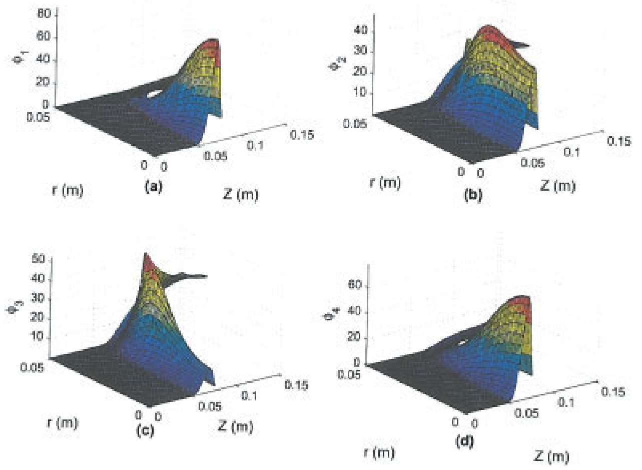

Figure 8. Dominant eigenfunctions for concentration of *MMGa* for inlet configurations (a) *TNN,* (b) *NTH,* (c) *TNT,* (d) *TNH,* respectively.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$$\left(\frac{\partial(\rho\mathbf{u})}{\partial t} + \nabla\cdot(\rho\mathbf{uu}) - \nabla\cdot\mathbf{T} - \rho\mathbf{g}, \psi_i^{u^r}\right) = 0, i = 1, \ldots, n_{u^r}$$

$$\left(\frac{\partial(\rho\mathbf{u})}{\partial t} + \nabla\cdot(\rho\mathbf{uu}) - \nabla\cdot\mathbf{T} - \rho\mathbf{g}, \psi_i^{u^z}\right) = 0, i = 1, \ldots, n_{u^z}$$

$$T = \sum_{i=1}^{n_T} a_i^T(t)\psi_i^T(r, z): \left(C_p\left[\frac{\partial(\rho T)}{\partial t} + \nabla\cdot(\rho\mathbf{u}T)\right] + \nabla\cdot(-k\nabla T), \psi_i^T\right) = 0, i = 1, \ldots, n_T$$

$$Y_k = \sum_{i=1}^{n_{Y_k}} a_i^{Y_k}(t)\psi_i^{Y_k}(r, z) : \left(\frac{\partial(\rho Y_k)}{\partial t} + \nabla\cdot(\rho\mathbf{u}Y_k) + \nabla\cdot\mathbf{j}_k - W_k\dot{\omega}_k, \psi_i^{Y_k}\right) = 0, k = 1, \ldots, N_s - 1, i = 1, \ldots, n_{Y_k}$$

$$t_f = [1 \;\; 1 \;\; 1 \;\; 1]\delta t, \; \delta t \geq 0 \qquad (22)$$

where $(\cdot, \cdot)$ denotes the inner product as defined in Eq. 7, *ls* represents species *TMGa, DMGa, MMGa* and *adduct; F* is the objective function, and R_{dep} is the surface deposition rate of GaN, $\bar{H}$ and H_{obj} are the spatially averaged thickness and the target thickness of the film at the end of deposition process, respectively. t_f represents the total process time and $\delta t = [\delta t_1 \; \delta t_2 \; \delta t_3 \; \delta t_4]$ is a four-dimensional vector representing the switching times. k_{ls} are the rate constants for surface reactions *S*1–*S*3. In the earlier optimization problem, R_o was taken to be 90% of the total wafer radius.[29] The rationale behind this approach is to avoid the edge effect, which was always present irrespective of the inlet configuration and operation. We will refer to R_o as cutoff radius whose numerical value was set to 1.35 in. The objective function *F* is a quadratic function that penalizes the spatial nonuniformity of the final film thickness across the wafer surface area within the cut-off

Table 4. Optimal Switching-Time Policy (Scheme 1)

Switching	Time [s]
HHH − NTH	0.00
NTH − TNN	2.43
TNN − TNH	3.75
TNH − TNT	5.27
TNT − HHH	6.81

radius, and the deviation of the spatially averaged film thickness from a predefined target value.

Finally, in the specific problem formulation, the time durations during which the reactor operates with a specific inlet configuration, δt_i, were the design variables. The nonlinear program was solved using a projected BFGS-Armijo algorithm[35] to compute optimal switching times from one inlet configuration to another. The gradients of the objective functional with respect to design variables were computed using finite differences. However, gradients can also be computed through the introduction of adjoint variables.

Remark 8: The optimization problem defined in Eq. 22 is a NLP with simple bound constraints, which was solved using a projected BFGS-Armijo algorithm. The algorithm employs structured quasi-Newton updating scheme for the Hessian, given by

$$H = C(x) + A,\ C(x) = \mathcal{P}_{\mathcal{A}^{\varepsilon(x)}}$$

$$A_{+}^{\dagger} = \left(I - \frac{sy^T}{y^T s}\right) A_c^{\dagger} \left(I - \frac{ys^T}{y^T s}\right) + \frac{ss^T}{y^T s}$$

$$(\mathcal{P}_{\mathcal{A}} + A)^{-1} = (\mathcal{P}_{\mathcal{A}} + A^{\dagger})$$

$$s = \mathcal{P}_{\mathcal{I}_+}(x_+ - x_c),\ y = \mathcal{P}_{\mathcal{I}_+}(\nabla f(x_+) - \nabla f(x_c)) \qquad (23)$$

where $\mathcal{I}_+ = \mathcal{I}^{\varepsilon+}(x_+)$, and $\mathcal{A}^{\varepsilon(x)}$ and $\mathcal{I}^{\varepsilon(x)}$ denote ε-active and ε-inactive sets at x, and the matrix $A^{\dagger}$ is the generalized inverse of A. The ε-active set at x is defined as

$$\mathcal{A}^{\varepsilon}(x) = \{i | U_i - (x)_i \leq \varepsilon \text{ or } (x)_i - L_i \leq \varepsilon\} \qquad (24)$$

and ε-inactive set is the compliment of $\mathcal{A}^{\varepsilon}(x)$. L and U are the lower and upper bounds on the optimization variable x. The operator $\mathcal{P}$ is defined for any set of indices, $\mathcal{S}$, as

$$(\mathcal{P}_{\mathcal{S}x})_i = \begin{cases} (x)_i, & \text{if } i \in \mathcal{S}, \\ 0, & \text{otherwise} \end{cases}$$

In case, if $y^T s \leq 0$, the algorithm reinitializes A to $\mathcal{S}_{\mathcal{I}}$. For further details of the algorithm and convergence analysis along with MATLAB implementation, refer to [35].

Optimization results

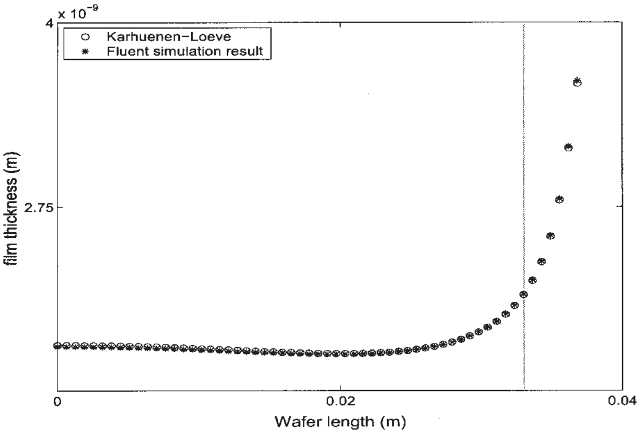

Figure 9. Final GaN film thickness computed through integration of reduced order model (stars) and full order model (circles).

Vertical line represents the cutoff radius.

In order to demonstrate the effectiveness of optimal switching of inlet configurations towards obtaining a final thin film of high radial uniformity, we considered a switching scheme comprising of switching from an *NTH* configuration to *TNN,* from *TNN* to *TNH,* and from *TNH* to *TNT* configurations. Initially the reactor was assumed to be operating at steady state with pure hydrogen flowing through the three inlets. A projected BFGS algorithm[35] was used to obtain the solution to the optimization problem given by Eq. 22; it took 52 searches and 558 s of CPU time to compute the solution on a Pentium 4 @ 2.4 GHz processor. The solution for optimal switching times is presented in Table 4. We note that the time needed for the computation of empirical eigenfunctions is not included in the calculation of the time needed to solve the optimization problem. The time needed to compute the empirical eigenfunctions was approximately 30 h, which is significantly lower than the estimated time required to solve the optimization problem when using the full-order model (520 h).

The thickness of the deposited film along the substrate obtained from integration of the reduced order model (for optimal switching of inlets) is shown in Figure 9, and is compared with results of Fluent simulations under the same policy. The error between the two is 1%. Thus, the use of a reduced-order model resulted in considerable savings of computational resources, with minimal loss of accuracy. In order to further demonstrate the accuracy of the reduced-order model,

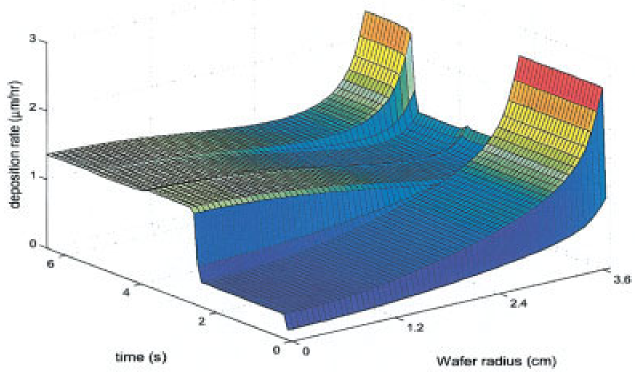

Figure 10. GaN deposition-rate profile across the substrate surface as a function of process time, calculated using FLUENT.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

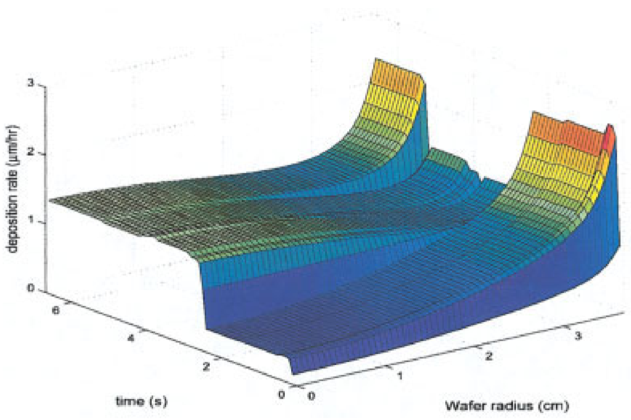

Figure 11. GaN deposition-rate profile across the substrate surface as a function of process time, computed using the reduced-order model.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

temporal variations of GaN deposition rate on the substrate are plotted in Figures 10 and 11 for simulations using fluent (full-order model) and empirical eigenfunctions (reduced-order model), respectively. It can be seen that the reduced-order model follows the full-order model closely for all times, and the error (shown in Figure 12) between the two is marginal.

In Figure 13, the final film thickness, at the end of the process operation, along the substrate radius is shown for each inlet configuration and compared against the optimal case. The extent of homogeneity achieved through switching is evident. Quantitatively, the maximum variation in wafer thickness from the center of the wafer is 86%, 32%, 35% and 20% for *NTH, TNN, TNH* and *TNT* inlet configurations, respectively, while for the optimal operation it is 3.1%. All profiles were obtained through Fluent simulations and the cutoff radius is represented by the dashed (green) line.

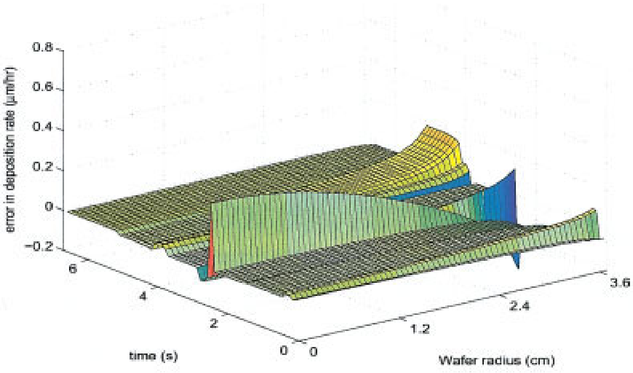

Figure 12. Spatiotemporal profile across the substrate surface of the error in GaN deposition-rate, using reduced-order and full-order models during optimal process operation.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

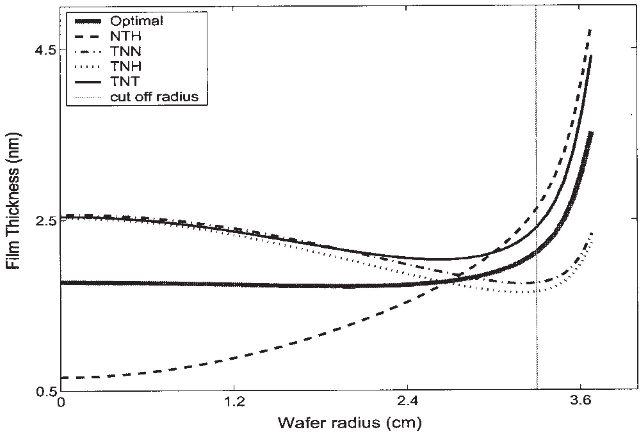

Figure 13. GaN film thickness at the end of process operation for NTH (dashed line), TNN (dash-dotted line), TNH (dotted line), and TNT (solid line) constant inlet configurations and under the optimal policy (thick solid line).

Cutoff radius is shown by vertical line.

Finally, in order to ascertain whether the optimal film thickness was dependent on the sequence with which inlet configurations were employed, an alternative switching scheme comprising of *HHH* → *TNH* → *NTH* → *TNT* → *TNN* was considered. Table 5 lists the corresponding optimal switching times for each inlet configuration and Figure 14 compares the final film deposited via switching schemes 1 and 2. It is evident that the resulting film thickness uniformity is only slightly affected by the switching sequence.

Conclusion

An approach that links nonlinear model reduction techniques with control vector parametrization-based schemes to efficiently solve (utilizing standard nonlinear programming techniques) dynamic constraint optimization problems arising in the context of spatially-distributed processes described by highly-dissipative nonlinear partial-differential equations was presented. Nonlinear model-reduction of the process description was accomplished through spatial discretization using empirical eigenfunctions and method of weighted residuals which takes advantage of the presence of low-order dominant dynamics in the solution of parabolic PDEs. The proposed approach was successfully applied to a MOVPE process for the production of *GaN* thin films, where it was demonstrated that the spatial non-uniformity (optimization objective) of the deposited film across the substrate surface can be reduced from 33% (steady-state operation with constant inlet *TNN* configuration) to 3% (under the optimal switching policy). It was

Table 5. Optimal Switching-Time Policy (Scheme 2)

Switching	Time [s]
HHH − TNH	0.00
TNH − NTH	1.45
NTH − TNT	3.78
TNT − TNN	5.24
TNN − HHH	6.70

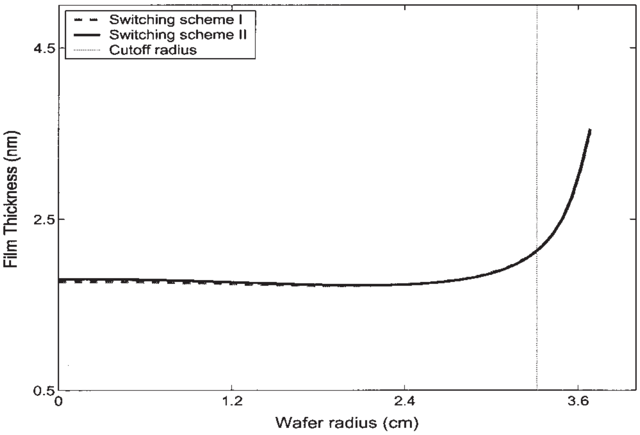

Figure 14. Comparison of GaN film thickness at the end of process operation employing switching policies of scheme 1 and 2.

Cutoff radius is shown by vertical lines.

also demonstrated that the resulting reduced-order NLP formulation is not only computationally efficient but also accurately predicts the process behavior, in order to accurately identify the optimal design variable profiles.

Acknowledgment

Financial support from the Dept. of Chemical Engineering at Pennsylvania State University is gratefully acknowledged.

Literature Cited

1. Vasantharajan S, Viswanathan J, Biegler LT. Reduced successive quadratic programming implementation for large-scale optimization problems with smaller degrees of freedom. *Comp & Chem Eng.* 1990;14:907–915.
2. Floudas CA, Panos MP. *Recent advances in global optimization.* Princeton, NJ: Princeton University Press; 1992.
3. Biegler LT, Nocedal J, Schmid C. Reduced Hessian strategies for large-scale nonlinear programming. *SIAM J of Optimization.* 1995;5:314.
4. Bendersky E, Christofides PD. A computationally efficient method for optimization of transport-reaction processes. *Comp & Chem Eng.* 1999;23(s):447–450.
5. Bendersky E, Christofides PD. Optimization of transport reaction processes using nonlinear model reduction. *Chem Eng Sci.* 2000;55: 4349–4366.
6. Armaou A, Christofides PD. Dynamic optimization of dissipative PDE systems using nonlinear order reduction. *Chem Eng Sci.* 2002;57: 5083–5114.
7. Lumley JL. Coherent Structures in Turbulence in *Transition and Turbulence.* Academic Press, New York, 1981;215–242.
8. Sirovich L. Turbulence and the dynamics of coherent structures: Part I: Coherent structures. *Quart Appl Math.* 1987;45:561–571.
9. Sirovich L. Turbulence and the dynamics of coherent structures: Part II: Symmetries and transformations. *Quart Appl Math.* 1987;45:573–582.
10. Christofides PD. *Nonlinear and Robust Control of PDE Systems: Methods and Applications to Transport-Reaction Processes.* Boston: Birkhäuser; 2001.
11. Temam R. *Infinite-Dimensional Dynamical Systems in Mechanics and Physics.* New York: Springer-Verlag; 1988.
12. Biegler LT, Cervantes AM, Wächter A. Advances in simultaneous strategies for dynamic process optimization. *Chem Eng Sci.* 2002;57: 575–593.
13. Binder T, Cruse A, Villar CAC, Marquardt W. Dynamic optimization using a wavelet based adaptive control vector parameterization strategy. *Comp & Chem Eng.* 2000;24:1201–1207.
14. Feehery WF, Barton PI. Dynamic optimization with equality path constraints. *Ind Eng Chem Res.* 1999;38:2350–2363.
15. Vassiliadis VS, Sargent RWH, Pantelides CC. Solution of a class of multistage dynamic optimization problems, Parts I & II. *Ind & Eng Chem Res.* 1994;33:2111–2133.
16. Serban R, Li ST, Petzold LR. Adaptive algorithms for optimal control of time-dependent partial differential-algebraic equation systems. *Int J Num Methods Eng.* 2003;57:1457–1469.
17. Balsa-Canto E, Banga JR, Alonso AA, Vassiliadis VS. Efficient optimal control of bioprocesses using second-order information. *Ind Eng Chem Res.* 2000;39:4287–4295.
18. Balsa-Canto E, Banga JR, Alonso AA, Vassiliadis VS. Dynamic optimization of distributed parameter systems using second-order directional derivatives. *Ind Eng Chem Res.* 2004;43:6756–6765.
19. Balsa-Canto E, Alonso AA, Banga JR. A novel, efficient and reliable method for thermal process design and optimization. Part I: Theory. *J of Food Eng.* 2002;52:227–234.
20. Balsa-Canto E, Banga JR, Alonso AA. A novel, efficient and reliable method for thermal process design and optimization. Part I: Applications. *J of Food Eng.* 2002;52:235–247.
21. Balsa-Canto E, Alonso AA, Banga JR. Reduced order models for nonlinear distributed process systems and their application in dynamic optimization. *Ind Eng Chem Res.* 2004;43:3353–3363.
22. Armaou A, Baker J, Christofides PD. Feedback control of plasma etching reactors for improved etching uniformity. *Chem Eng Sci.* 2001;56:257–265.
23. Armaou A, Christofides PD. Plasma enhanced chemical vapor deposition: Modeling and control. *Chem Eng Sci.* 1999;54:3305–3314.
24. Theodoropoulou A, Zafiriou E, Adomaitis RA. Inverse model-based real-time control for temperature uniformity of RTCVD. *IEEE Transactions on Semiconductor Manufacturing.* 1999;12:87–101.
25. Theodoropoulou A, Adomaitis RA, Zafiriou E. Model reduction for optimization of rapid thermal chemical vapor deposition. *IEEE Transactions on Semiconductor Manufacturing.* 1998;11:85–98.
26. Baker J, Christofides PD. Finite dimensional approximation and control of nonlinear parabolic PDE systems. *Int J Contr.* 2000;73:439–456.
27. Itle GC, Salinger AG, Pawlowski RP, Shadid JN, Biegler LT. A tailored optimization strategy for PDE-based design: Application to a CVD reactor. *Comp & Chem Eng.* 2004;28:291–302.
28. Theodoropoulos C, Ingle NK, Mountziaris TJ. Computational studies of the transient behavior of horizontal MOVPE reactors. *J Crystal Growth.* 1997;170:72–76.
29. Theodoropoulos C, Mountziaris TJ, Moffat HK, Han J. Design of gas inlets for the growth of gallium nitride by metalorganic vapor phase epitaxy. *J Crystal Growth.* 2000;217:65–81.
30. Polak E. On the mathematical foundations of nondifferentiable optimization in engineering design. *SIAM Review.* 1987;29:21–89.
31. Hettich R, Kortanek KO. Semi-infinite programming: Theory, methods, and applications. *SIAM Review.* 1993;35:380–429.
32. Graham MD, Kevrekidis IG. Alternative approaches to the Karhunen-Loève decomposition for model reduction and data analysis. *Comp & Chem Eng.* 1996;20:495–506.
33. Holmes P, Lumley JL, Berkooz G. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry.* New York: Cambridge University Press; 1996.
34. Fukunaga K. *Introduction to Statistical Pattern Recognition.* New York: Academic Press; 1990.
35. Kelley CT. *Iterative Methods for Optimization; 18 of Frontiers in Applied Mathematics.* Philadelphia, PA, USA: SIAM 1999. MATLAB subroutines available at http://www4.ncsu.edu/˜ ctk/matlab_darts.html.
36. Luus R, Jaakola THI. Optimization by direct search and systematic reduction of the size of search region. *AIChE J.* 1973;19:760–766.
37. Nakamura S, Fasol G. *The Blue Laser Diode.* Berlin, Heidelberg: Springer; 1997.
38. Amano H, Sawaki N, Akasaki N, Toyoda Y. Metalorganic vapor phase epitaxial growth of a high quality GaN film using an AlN buffer layer. *Appl Phys Lett.* 1986;48:353–356.
39. Wuu DS, Horng RH, Tseng WH, Lin WT, Kung CY. Influences of temperature ramping rate on GaN buffer layers and subsequent GaN overlayers grown by metalorganic chemical vapor deposition. *J Crystal Growth.* 2000;220:235–242.
40. Kisker DW, Kuech TF. *MOCVD Technology and Its Applications,* In:

DTJ Hurle, *Handbook on Semiconductors,* Vol. 3. Amsterdam: Elsevier Science; 1994.

41. Yang CC, Chuag-Kuei H, Chi GC, Chyi WM. Growth and characterization of GaN by atmospheric pressure metalorganic chemical-vapor deposition with a novel separate-flow reactor. *J Crystal Growth.* 1999; 200:39–44.
42. Kuech TF, Gu S, Wate R, Zhang L, Sun J, Dumesic JA, Redwing JM. The chemistry of GaN growth. In: *Materials Research Society Symposium Proceedings.* 2001;639:G1.1.1–11.
43. Safvi SA, Redwing JM, Tischler MA, Kuech TF. GaN growth by metallorganic vapor phase epitaxy: A comparison of modeling and experimental measurements. *J Electrochem Soc.* 1997;144:1789–1796.
44. Safvi SA, Redwing JM, Thon A, Flynn JS, Tischler MA, Kuech TF. MOVPE GaN gas phase chemistry for reactor design and optimization. *Mat Res Soc Symp Proc.* 1997;449:101–106.
45. Sengupta D. Does the ring compound $[(CH_3)_2Ga : NH_2]_3$ form during MOVPE of gallium nitride? Investigations via density functional theory. *Comp & Chem Eng.* 2004;28:291–302.

Manuscript received Oct. 19, 2004, and revision received Aug. 5, 2005.