Dynamic Optimization of Dissipative PDEs Using Control Vector Parameterization: Application to GaN Thin Film Epitaxy*

Antonios Armaou† and Amit Varshney
Department of Chemical Engineering
Pennsylvania State University, University Park, PA 16802

***Abstract*— We present an approach that links nonlinear model reduction techniques with control vector parametrization-based schemes to efficiently solve** $dynamic$ **constraint optimization problems arising in the context of spatially-distributed processes governed by highly-dissipative nonlinear partial differential equations (PDEs). The proposed approach is applied to a Metal-Organic Vapor-Phase Epitaxy process for the production of GaN thin films, with the objective to minimize the spatial non-uniformity of the deposited film across the substrate surface.**

I. Introduction

The traditional assumption towards the design of industrially important chemical processes has been the consideration of steady-state operating conditions. However, there are instances where more efficient process operation can be accomplished through time-varying operation. The increase of available computational power has prompted the use of direct search methods for the computation of optimal, with respect to a specific cost function, time-varying operating conditions.

To address the issue of optimal operation for dynamically evolving processes, computational approaches have been proposed that may involve the discretization of both the control and process variables in the temporal domain, solved using sparse linear algebra techniques [24], [5], [3], [10], [6], or formulations necessitating direct integration of the model equations in time, keeping track of possible parameter and path constraint violations and temporal discretization of the operating variables [5], [7], [9], [25], [20]; the control vector parametrization method.

Contrary to the wealth of results on optimization of dynamically evolving lumped parameter processes, only recently the computation of optimal strategies for processes that involve coupling of complex chemical reactions with significant mass and energy transport mechanisms was addressed [4], [2]. Mathematical descriptions of transport-reaction processes can be derived from dynamic conservation equations and usually involve highly dissipative (typically parabolic) partial differential equation (PDE) systems that hinder the use of standard search algorithms due to the large number of ordinary differential equations (ODEs) in the formulation, necessary to accurately capture the spatial variation of the process. Motivated by this, computationally tractable programs for spatially distributed processes at steady state were recently formulated [4], using nonlinear model reduction techniques. In [2], a computationally efficient procedure was developed for dynamically evolving highly dissipative systems, using nonlinear model reduction techniques to map them as low-order differential-algebraic equations (DAEs).

In this work, we present an approach that links nonlinear model reduction techniques with control vector parametrization-based schemes to efficiently solve $dynamic$ constraint optimization problems arising in the context of spatially-distributed processes governed by highly-dissipative nonlinear partial differential equations (PDEs). The approach is based on combination of the method of weighted residuals with spatially-global empirical eigenfunctions as basis functions (constructed by applying Karhunen-Loève expansion [16], [21] to appropriately composed ensembles of process data) to spatially discretize the PDEs and derive greatly reduced in order, yet highly accurate, ODE models. We apply the proposed approach to achieve uniform radial thickness of GaN films in a vertical MOVPE reactor. 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 solve a constraint dynamic optimization problem that determines the optimal precursor concentration spatiotemporal profiles at the inlet to grow films which have a high degree of spatial uniformity.

II. Preliminaries

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

$$\begin{aligned} \frac{\partial x}{\partial t} &= \mathcal{A}(x) + f(t,x,d), \quad x(z,0) = x_0(z) \\ g(x, \frac{dx}{d\eta}, \ldots, \frac{d^{n_o-1}x}{d\eta^{n_o-1}}) &= 0, \quad on\ \Gamma \end{aligned} \tag{1}$$

*Financial support for this work from the Pennsylvania State University, Chemical Engineering Department, is gratefully acknowledged.

†Corresponding author armaou@psu.edu

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. $\mathcal{A}(x)$ 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, $g(x, \frac{dx}{d\eta}, \ldots, \frac{d^{n_o-1}x}{d\eta^{n_o-1}})$ is a nonlinear vector function which is assumed to be sufficiently smooth (n_o is the order of the PDE of Eq.1), $\left.\frac{dx}{d\eta}\right|_\Gamma$ 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 is applicable to a wide range of dynamic spatially distributed processes including both transport-reaction processes and several classes of dissipative fluid dynamic systems [8]. The nonlinear structure of the spatial differential operator, $\mathcal{A}(x)$, 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 present in section VI.

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

$$\begin{aligned} &min \int_0^{t_f} \int_\Omega G(x(z,t), d(t)) dz dt \\ &\quad s.t \\ &-\frac{\partial x}{\partial t} + \mathcal{A}(x) + f(t,x,d) = 0, \\ &x(z,0) = x_0(z), \;\; g(x, \frac{dx}{d\eta}, \ldots, \frac{d^{n_o-1}x}{d\eta^{n_o-1}}) = 0 \;\; on \; \Gamma \\ &g(x,d) \leq 0, \;\; \forall \, z \in \Omega, \;\; t \in [0, t_f] \end{aligned} \tag{2}$$

where $\int_0^{t_f} \int_\Omega G(x,d) dz dt$ is the objective function 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.

III. Spatial discretization

A. 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:

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

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, where the optimization parameters are the design variables $d(t)$ and the time varying coefficients $a_{kN}(t)$:

$$\begin{aligned} &min \int_0^{t_f} \int_\Omega G(\sum_{k=1}^{N} a_{kN}(t)\phi_k(z), d) dz dt \\ &s.t \\ &-\sum_{k=1}^{N} \dot{a}_{kN} (\int_\Omega \psi_\nu(z)\phi_k(z) dz) \\ &+ \int_\Omega \psi_\nu(z) \mathcal{A}(\sum_{k=1}^{N} a_{kN}(t)\phi_k(z)) dz \\ &+ \int_\Omega \psi_\nu(z) f(t, \sum_{k=1}^{N} a_{kN}(t)\phi_k(z), d) dz = 0 \\ &\int_\Omega \psi_\nu(z) g(\sum_{k=1}^{N} a_{kN}\phi_k(z), d) dz \leq 0 \end{aligned} \tag{4}$$

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 equality and inequality depends on the choice of the weighting functions, as well as on N. Owing to the smoothness of the functions $G(x,d)$, $\mathcal{A}(x)$, $f(t,x,d)$, $g(x,d)$ 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 $N \to \infty$.

B. Computation of empirical eigenfunctions via Karhunen-Loève expansion

To perform spatial discretization of the nonlinear program of Eq.2, we use solution data of the system of Eq.1 to construct global basis functions via Karhunen-Loève (K-L) expansion. The motivation for following this approach is provided by the occurrence of dominant spatial patterns in the solution of several dissipative PDEs and the presence of actuation (external influence) and the need to account for those in the shape of the global basis functions. 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 [12], [4], [2] for more 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 values of d_m in each one of those intervals are denoted by $d_{m,j}$, $j = 1, \ldots, m_{d_m}$.
- 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, say $d_{m,j}$, and subsequently computing $d_m(t)$ for the entire time interval of process operation using linear interpolation.
- Finally, we compute a set of PDE solution data (ensemble) for all possible combinations of initial conditions and profiles of $d(t)$.

Application of K-L expansion to the 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 (empirical eigenfunctions); the reader may refer to [11], [13], [2] for a detailed presentation of the method. 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. 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 any number of modes L. Therefore, the K-L expansion yields the most efficient way for computing the basis functions (corresponding to the largest empirical eigenvalues) capturing the dominant patterns of the ensemble.

Remark 1: We note that the basis that we compute using K-L decomposition is independent of the functional that we try to minimize. Therefore, the same basis can be used to perform computationally efficient optimizations with respect to different functionals associated with the same underlying set of partial differential equations.

Remark 2: We note that 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 3: As a practical implementation note, we point out that 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.

IV. Temporal Discretization

The computational solution of semi-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 number of variables for the subsequent numerical solution. Specifically, we use control vector parametrization (CVP) scheme to reformulate the dynamic program of Eq.4 as an algebraic nonlinear one. CVP (also known as shooting method) involves the temporal discretization of the control parameter vector only, and the solution of the dynamic equality constrains through direct integration, keeping track of constraint violations during the process evolution [25], [9].

We discretize the temporal domain into m_t intervals and define the temporal discretization step as $\delta t_i = t_i - t_{i-1}, \ \forall i = 1, ..., m_t$. The vector function $d(t)$ is then expressed as a series of the form

$$d(t) = \sum_{i=0}^{m_t - 1} d_{i+1}[H(t - t_i) - H(t_{i+1} - t)] \qquad (5)$$

where $H(\cdot)$ is the standard Heaviside function. Applying the above approximation to the dynamic nonlinear program of Eq.4, we obtain an algebraic nonlinear program of dimension $p \times m_t + N$, which has the following general form:

$$\begin{array}{c} minF(x) \\ s.t \\ h(x) = 0 \\ g(x) \leq 0 \end{array} \qquad (6)$$

where the explicit form of the functions $F(x), h(x), g(x)$ is omitted for brevity. Note that in the above formulation the integrated ODEs appear in $h(x)$.

V. 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 [14], [17]. 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 (P) in following algorithm:

- **Step 1**: Compute an initial guess for N, say $\hat{N}$, based on the magnitude of the eigenvalues corresponding to the eigenfunctions.
- **Step 2**: Use the spatial and temporal discretization procedures of sections 3 and 4, respectively, to derive a finite-dimensional program of the form of Eq.6.

- **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.6 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), 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 above 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 are not important). This is a consequence of the hierarchy of the eigenfunctions. On the other hand, the convergence of the above algorithm is a direct consequence of the fact that as N increases and δt decreases, the finite-dimensional program converges to the infinite-dimensional one.

VI. GaN thin film epitaxy: Process description and optimization problem formulation

In an economic environment where profit margins tighten due to market saturation, company survival dictates reduction of cost; thus the objective for microelectronics fabrication processes is to increase yield while minimizing reactant consumption and satisfying safety requirements. One such process is the metal-organic vapor phase epitaxy (MOVPE), also known as metal-organic chemical vapor deposition; MOVPE is the 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.

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 depends 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 [22]. 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 mass transport of the precursor gases, adsorption, subsequent surface diffusion and reaction of the adsorbed species, and desorption.

GaN is one such semiconductor and because of its wide bandgap energy (3.4 eV), it has potential applications in manufacture of blue-green LED and laser diodes. Currently GaN on Sapphire, Si or SiC substrates is produced in a two stage MOVPE process [18], [1] usually with trimethygallium ($TMGa$) and NH_3 as precursors for gallium and nitrogen respectively, diluted in H_2 (carrier gas). During the first stage, a GaN nucleation layer is formed on the wafer surface at low temperatures (600 $^\circ C$), 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. Alternatively, AlN is also employed as a buffer layer [1]. 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 of 40 $^\circ C/min$ to be optimal in terms of the resulting GaN layer quality [26]. The second stage of the GaN epitaxy is initiated at temperature of $1060\ ^\circ C$, forming the desired GaN epilayer.

Parasitic pre-reactions between Ga and N_2 precursors that form Lewis acid-base adducts are known to occur, which on one hand deplete the feed stream 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. One way to avoid this problem is to feed gallium and nitrogen containing precursors from different inlets, so that mixing between the two occurs just above the wafer, rather than using a single inlet of premixed stream of precursors.

In order to grow films of uniform thickness a spatially invariant concentration of Ga containing precursors over the substrate is necessary, a requirement which is impossible to meet because of transport and reaction limitations. Recent simulation results [23] 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 number of inlets does not guarantee a high degree of thickness uniformity.

In this paper, we demonstrate an alternate approach to achieve uniform radial thickness of GaN films in a vertical MOVPE reactor. The approach is 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 propose that by appropriately switching from one deposition rate profile to another, it is possible to grow films which have a high degree of spatial uniformity.

VII. Process model derivation

A schematic of vertical MOVPE reactor with showerhead configuration is shown in Figure 1. Precursor gases for gallium and nitrogen enter through the inlet directly above the substrate over which the film is deposited. $TMGa$ and ammonia diluted in hydrogen carrier gas were used as precursors for gallium and nitrogen respectively. In order to obtain spatial variations in concentration of precursors across the inlet, a split inlet design comprising of a three

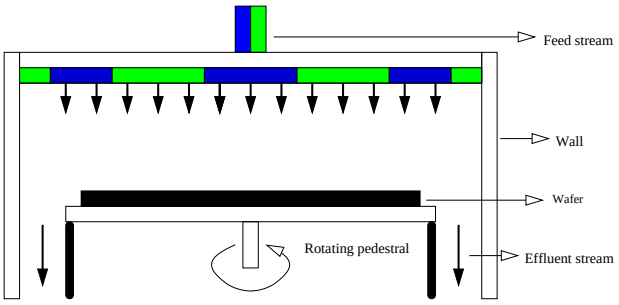

Fig. 1. vertical MOVPE reactor with a three concentric ring showerhead inlet configuration .

concentric ring showerhead reactor was used. Further details of process conditions and reactor geometry are provided in Table I.

TABLE I

REACTOR GEOMETRY AND PROCESS CONDITIONS

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}
$X^*_{NH_3}$		0.15

* Inlet Mole fractions of reactant in H_2 carrier gas

The reaction model describing the reactions between gas-phase species and gas-surface reactions has been adopted from [23], [15] and is shown is Table II. 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 and its rate is estimated by the rate of bimolecular collisions [23], 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 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. Formation of cyclic trimer has been shown to be negligible and thus is neglected from the gas-phase kinetic model [19].

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 our simulations reveal that their contribution towards the overall film growth is negligible.

The above reaction model was incorporated into momentum, energy and mass conservation equations, and the following set of coupled partial differential equations was solved using FLUENT.

$$\begin{aligned}
\frac{\partial \rho}{\partial t} + \bigtriangledown \cdot (\rho \mathbf{u}) &= 0 \\
\frac{\partial (\rho \mathbf{u})}{\partial t} + \bigtriangledown \cdot (\rho \mathbf{u}\ \mathbf{u}) - \bigtriangledown \cdot \mathbf{T} - \rho \mathbf{g} &= 0 \\
C_p[\frac{\partial (\rho T)}{\partial t} + \bigtriangledown \cdot (\rho \mathbf{u} T)] = - \bigtriangledown \cdot \mathbf{q} - \sum_k h_k W_k \dot{\omega} \qquad &(7) \\
\frac{\partial (\rho Y_k)}{\partial t} + \bigtriangledown \cdot (\rho \mathbf{u} Y_k) = - \bigtriangledown \cdot \mathbf{j}_k + W_k \dot{\omega}_k, & \\
\forall\ k = 1, \ldots, N_s - 1 &
\end{aligned}$$

where ρ is the density and C_p is the specific heat capacity of the multicomponent mixture, $\mathbf{u}$ is the fluid velocity vector and T is the temperature, $\mathbf{T}$ is the stress tensor and $\mathbf{q}$ is the heat flux due to conduction. Y_k is the mass fraction, h_k the partial specific enthalpy, W_k the molecular weight, and $\dot{\omega}_k$ the net production rate, due to homogeneous reactions, of species k. N_s is the number of gaseous species and $\mathbf{j}_k$ the species mass fluxes.

Physical properties 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. Full multicomponent diffusion model was used for species diffusion.

VIII. REDUCED-ORDER MODELLING AND OPTIMIZATION

Based on the results of Fluent simulations the following simplifying assumptions were made while formulating the optimization problem. It was found that the concentrations of species CH_4 and $(CH_3)_2Ga : NH_2$ were insignificant in comparison to concentration of other gaseous species inside the reactor. Hence, these species and the corresponding gas-phase and surface reactions ($G5$, $S5$ and $S6$) were omitted. Similarly, variations in axial and radial velocities were small irrespective of the inlet configuration. Thus, axial and radial velocities were assumed to be time invariant and equal to their respective time averages. Furthermore, heat generation due to chemical reactions was ignored because of low concentration of reacting species (e.g., $TMGa$).

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 drop the convective heat transfer term from the energy conservation equation. Also the dependence of specific heat and thermal conductivity of the mixture on temperature and mixture composition

TABLE II

GAS AND SURFACE REACTIONS

Gas phase Reactions		k0	E	Surface reactions†		$s^{\ddagger}$
(G1)	$Ga(CH_3)_3 \rightarrow Ga(CH_3)_2 + CH_3$	3.5×10^{15}	59.5	(S1)	$Ga(CH_3)_3 + S \rightarrow Ga(bulk) + 3CH_3$	1
(G2)	$Ga(CH_3)_2 \rightarrow GaCH_3 + CH_3$	8.7×10^{7}	35.4	(S2)	$Ga(CH_3)_2 + S \rightarrow Ga(bulk) + 2CH_3$	1
(G3)	$Ga(CH_3)_3 + NH_3 \rightarrow (CH_3)_3Ga:NH_3$	coll.	0	(S3)	$GaCH_3 + S \rightarrow Ga(bulk) + CH_3$	1
(G4)	$(CH_3)_3Ga:NH_3 \rightarrow Ga(CH_3)_3 + NH_3$	1×10^{14}	18.5	(S4)	$NH_3 + S \rightarrow N(bulk) + CH_3$	$-^{*}$
(G5)	$(CH_3)_3Ga:NH_3 \rightarrow (CH_3)_2Ga:NH_2 + CH_4$	1×10^{14}	49	(S5)	$(CH_3)_3Ga:NH_3 + 2S \rightarrow GaN + 3CH_4$	1
				(S6)	$(CH3)_2Ga:NH_2 + 2S \rightarrow GaN + 2CH_4$	1

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

was found to be almost the same. These considerations allowed the energy equation to be decoupled from the rest of the equations and an exponentially decaying relationship for deviation of temperature from the steady-state profile was assumed, whose time constant was tuned to minimize the error. 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. The above argument was further bolstered by the presence of a single dominant eigenvalue (which captured 98% of the energy included in the ensemble of snapshots) for temperature during calculation of the empirical eigenfunctions via Karhunen-Loève expansion.

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 non-uniformity in the final film thickness, the goal of optimization is to ascertain an optimal switching policy for inlet configurations.

A reduced-order model was obtained for the solution of the optimization problem through spatial discretization using the method of weighted residuals with empirical eigenfunctions (obtained by Karhunen-Loève expansion) as basis functions. 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 describing momentum, energy and mass transport. 28 snapshots were taken from each switching and a total of 25 different switchings (from 6 different inlet configurations) were employed to generate an ensemble of 28×25 snapshots. Under the assumptions discussed earlier, we required the computation of empirical eigenfunctions for five species namely $TMGa$, $DMGa$, $MMGa$, $adduct$ and NH_3 to derive the reduced-order model. In consequence, the derived reduced-order model based on the computed empirical eigenfunctions, involved 64 ODEs to describe the dynamic behavior of these species, with axial and radial velocities held constant at their respective ensemble averages and an algebraic expression describing the exponential (decaying) dependence of the spatial variations of temperature from the respective steady state.

The optimization problem was formulated as:

$$\begin{aligned} minF =& w_1 \int_0^{R_o} \left\{ \int_0^{t_f} (R_{dep}(r,t)dt - \bar{H}(\delta t))dt \right\}^2 dr \\ &+ w_2[H_{obj} - \bar{H}(\delta t)]^2 \\ & \text{s.t.} \\ & \bar{H} = \frac{1}{R_o} \int_0^{R_o} \int_0^{t_f} R_{dep}(r,t)dtdr \\ & t_f = [1\ 1\ 1\ 1]\delta t, \quad \delta t \geq 0 \\ & R_{dep}(r,t) = \textstyle\sum_{ls} k_{ls}(T_s) C_{ls}(t,r,z=z_0) \end{aligned} \tag{8}$$

where 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 average thickness and the target thickness of the film at the end of deposition process respectively. t_f represents the total process time and $\delta t = [\delta t\ \delta t_1\ \delta t_2\ \delta t_3\ \delta t_4]$ is a four-dimensional vector representing the switching times. k_{ls} are the rate constants for surface reactions $S1$-$S3$. In the above optimization problem, R_o was taken to be 90 percent of the total wafer radius [23]. The rationale behind this approach is to avoid the edge effect, which was always present irrespective of the inlet configuration. Throughout the rest of this paper, we will refer to it as cutoff radius. 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 radius and the deviation of the spatially averaged film thickness from a predefined target one.

Additional constraints to the optimization problem (in addition to the ones of Eq.8) arise from the Karhunen-Loève expansion, and are of the form of Eq.4. Their explicit form is omitted for brevity.

In the specific problem formulation the time duration of each of the available inlet configurations, δt_i, were the design variables. The non-linear program was solved using a projected BFGS algorithm [14] to compute optimal switching times from one inlet configuration to another.

IX. RESULTS

To better present the optimization results based on the inlet concentration profiles, we assign to each inlet configuration a shorthand notation; for example, we refer to an inlet configuration with NH_3 precursor flowing in the innermost inlet and $TMGa$ precursor flowing in the middle (with

H_2 carrier gas for both inlets) and pure H_2 flowing in the outermost inlet as NTH configuration; shorthand notations for other configurations are based on similar lines. 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 total of 15 empirical eigenfunctions were used for the five gaseous species (which accounted for at least 99 percent of the energy embedded in the ensemble of snapshots) with a different set of eigenfunctions (and in extension a different ODE model) for each switching. A projected BFGS algorithm [14] was used to obtain the solution to the optimization problem, which took 52 searches and 558 seconds of CPU time to reach at the solution. The solution for optimal switching times is presented in Table III. 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 hours, which is still lower than the estimated time required to solve the optimization problem when using the full-order model (520 hours).

TABLE III
OPTIMAL SWITCHING TIME

Switching	time [s]
HHH - NTH	0.00
NTH - TNN	2.40
TNN - TNH	3.70
TNH - TNT	5.20
TNT - HHH	6.70

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 2, and is compared with results of Fluent simulations (under the same policy). The green line represents the cutoff radius. The error between the two is 1 %. It took more than 10 hours to run one dynamic simulation in Fluent. Thus, the use of reduced order model resulted in considerable saving of computational resources, with minimal loss of accuracy. In order to further demonstrate the accuracy of the reduced-order model, temporal variations of GaN deposition rate on the substrate are plotted in Figures 3 and 4 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 (Figure 5) between the two is marginal.

In Figure 6, the final film thickness, at the end of the process operation $t_f = 6.7\ s$, 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

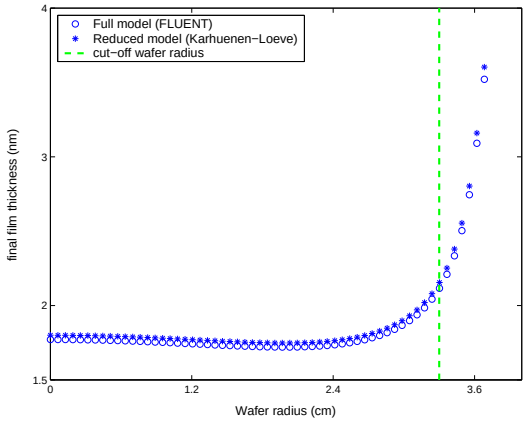

Fig. 2. Final GaN film thickness computed through integration of reduced-order model (stars) and from full order model (circles). Green line represents the cutoff radius.

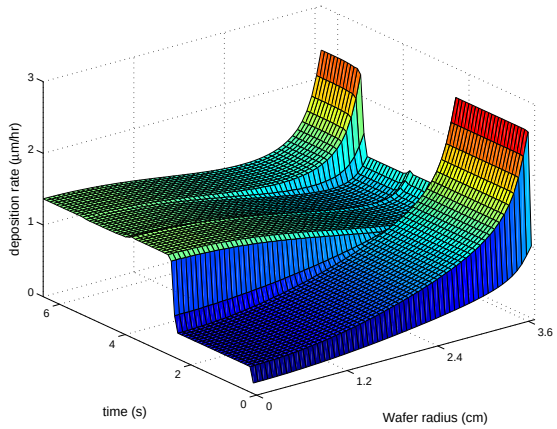

Fig. 3. GaN deposition rate on the heated substrate as a function of process time calculated using FLUENT.

variation in wafer thickness from 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 green (dashed) line.

X. CONCLUSION

An approach that linked non-linear model reduction techniques with control vector parameterization-based schemes to efficiently solve dynamic constraint optimization problems arising in the context of spatially-distributed processes described by highly-dissipative nonlinear partial differential equations was presented. The proposed approach was successfully applied to a MOVPE process, 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).

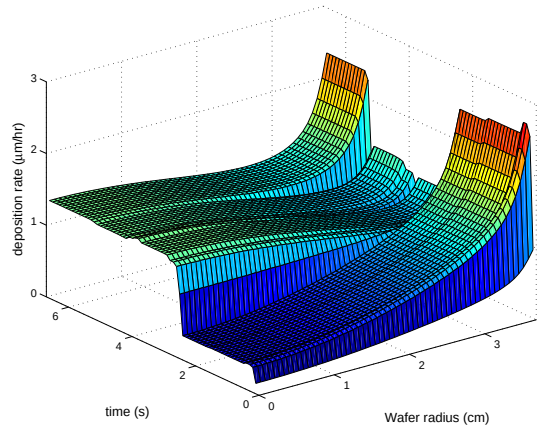

Fig. 4. GaN deposition rate on the heated substrate as a function of process time for the reduced-order model.

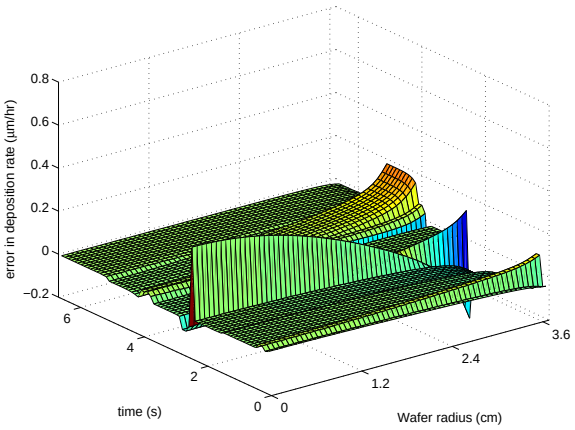

Fig. 5. Difference between GaN deposition rate computed using full-order and reduced-order models.

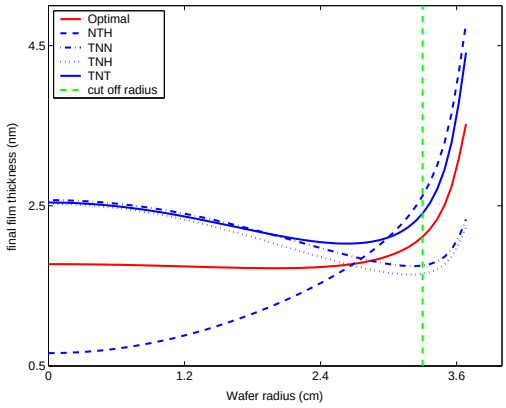

Fig. 6. GaN film thickness at the end of process operation for NTH (dashed line), TNN (dash-dot line), TNH (dotted line) and TNT (solid blue line) constant inlet configurations and under the optimal policy (solid red line). Green line represents the cut off radius.

REFERENCES

[1] H. Amano, N Sawaki, I. Akasaki, and Y. Toyoda. Metalorganic vapor phase epitaxial growth of a high quality GaN film using an AlN buffer layer. *Appl. Phys. Lett.*, 48:353–356, 1986.

[2] A. Armaou and P. D. Christofides. Dynamic optimization of dissipative PDE systems using nonlinear order reduction. *Chem. Eng. Sci.*, 57:5083–5114, 2002.

[3] M. P. Avraam, N. Shah, and C. C. Pantelides. A decomposition algorithm for the optimisation of hybrid dynamic processes. *Comp. & Chem. Engng.*, 23:S451–S454, 1999.

[4] E. Bendersky and P. D. Christofides. Optimization of transport-reaction processes using nonlinear model reduction. *Chem. Eng. Sci.*, 55:4349–4366, 2000.

[5] L. T. Biegler, A. M. Cervantes, and A. Wächter. Advances in simultaneous strategies for dynamic process optimization. *Chem. Eng. Sci.*, 57:575–593, 2002.

[6] L. T. Biegler, J. Nocedal, and C. Schmid. Reduced hessian strategies for large-scale nonlinear programming. *SIAM Journal of Optimization*, 5:314, 1995.

[7] T. Binder, A. Cruse, C. A. C. Villar, and W. Marquardt. Dynamic optimization using a wavelet based adaptive control vector parameterization strategy. *Comput. Chem. Eng.*, 24:1201–1207, 2000.

[8] P. D. Christofides. *Nonlinear and Robust Control of PDE Systems: Methods and Applications to Transport-Reaction Processes*. Birkhäuser, Boston, 2001.

[9] W. F. Feehery and P. I. Barton. Dynamic optimization with equality path constraints. *Ind. Eng. Chem. Res.*, 38:2350–2363, 1999.

[10] C. A. Floudas and M. P. Panos. *Recent advances in global optimization*. Princeton University Press, Princeton, N.J., 1992.

[11] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, New York, 1990.

[12] M. D. Graham and I. G. Kevrekidis. Alternative approaches to the Karhunen-Loève decomposition for model reduction and data analysis. *Comp. & Chem. Eng.*, 20:495–506, 1996.

[13] P. Holmes, J. L. Lumley, and G. Berkooz. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press, New York, 1996.

[14] C. T. Kelley. *Iterative Methods for Optimization*, volume 18 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, PA, USA, 1999.

[15] T. F. Kuech, S. Gu, R. Wate, L. Zhang, J. Sun, J.A. Dumesic, and J.M. Redwing. The chemistry of GaN growth. In *Materials Research Society Symposium Proceedings*, volume 639, pages G1.1.1–11, 2001.

[16] J. L. Lumley. Coherent structures in turbulence. In *Transition and Turbulence*, pages 215–242, Academic Press, New York, 1981.

[17] R. Luus and T. H. I. Jaakola. Optimization by direct search and systematic reduction of the size of search region. *AIChE J.*, 19:760–766, 1973.

[18] S. Nakamura and G. Fasol. *The Blue Laser diode*. Springer, Berlin, Heidelberg, 1997.

[19] D. Sengupta. Does the ring compound $[CH_3)_2GaNH_2]_3$ form during MOVPE of gallium nitride? investigations via density functional and reaction rate theories.

[20] R. Serban, S. T. Li, and L. R. Petzold. Adaptive algorithms for optimal control of time-dependent partial differential-algebraic equation systems. *Int. J. Num. Methods Engng.*, 57:1457–1469, 2003.

[21] L. Sirovich. Turbulence and the dynamics of coherent structures: part I: Coherent structures. *Quart. Appl. Math.*, XLV:561–571, 1987.

[22] C. Theodoropoulos, N.K. Ingle, and T.J. Mountziaris. Computational studies of the transient behavior of horizontal movpe reactors. *J. Crystal Growth*, 170:72–76, 1997.

[23] C. Theodoropoulos, T. J. Mountziaris, H. K. Moffat, and J. Han. Design of gas inlets for the growth of gallium nitride by metalorganic vapor phase epitaxy. *J. Crystal Growth*, 217:65–81, 2000.

[24] S. Vasantharajan, J. Viswanathan, and L. T. Biegler. Reduced successive quadratic programming implementation for large-scale optimization problems with smaller degrees of freedom. *Computers Chem. Eng.*, 14:907–915, 1990.

[25] V. S. Vassiliadis, R. W. H. Sargent, and C. C. Pantelides. Solution of a class of multistage dynamic optimization problems, parts I & II. *Ind. & Eng. Chem. Res.*, 33:2111–2133, 1994.

[26] D.-S. Wuu, R.-H. Horng, W.-H. Tseng, W.-T. Lin, and C.-Y. Kung. Influences of temperature ramping rate on GaN buffer layers and subsequent GaN overlayers grown by metalorganic chemical vapor deposition. *J. Crystal Growth*, 220:235–242, 2000.