# RECENT RESULTS IN SOLVING INDEX 2 DIFFERENTIAL-ALGEBRAIC EQUATIONS IN CIRCUIT SIMULATION

ROSWITHA MÄRZ AND CAREN TISCHENDORF*

dedicated to C. W. Gear on the occasion of his 60th anniversary

**Abstract.** In electric circuit simulation the charge oriented modified nodal analysis may lead to highly nonlinear DAEs with low smoothness properties. They may have index 2 but they do not belong to the class of Hessenberg form systems that are well understood.

In the present paper, on the background of a detailed analysis of the resulting structure, it is shown that charge oriented modified nodal analysis yields the same index as the classical modified nodal analysis.

Moreover, for index 2 DAEs in the charge oriented case, a further careful analysis with respect to solvability, linearization and numerical integration is given.

**Key words.** Differential-algebraic equations, index 2, circuit simulation, IVP, numerical integration, BDF, defect correction.

**AMS subject classifications.** 65L10

**1. Introduction.** In modern circuit simulation, the so-called charge oriented modified nodal analysis is preferred for different reasons ([4], [7]). The resulting DAEs have low smoothness properties. They may have index 2, but they do not have Hessenberg form at all.

In the Sections 2 and 3 of the present paper, by investigating the structural conditions in more detail, it is shown that both the classical modified nodal analysis and the charge oriented modified nodal analysis lead to DAEs of the same tractability index. Furthermore, the constant leading nullspace seems to be an advantage of the charge oriented formulation.

Moreover, a further analysis of index-2 DAEs resulting from modified nodal analysis is given in Section 4. The solvability of initial value problems is stated under low smoothness. It is shown how the solutions depend on the initial data. The sensitivity matrix satisfies again the linearized system. In particular, certain relevant projectors and subspaces are described in detail.

In Section 5 we discuss the behaviour of the BDF applied to DAEs of the class under consideration. A more general result from [20] on weak instability is specified on the background of the special structure given in Section 4. In particular, also the error propagation due to the weak instability is considered. Unfortunately, in nonlinear DAEs all solution components may be affected whereas in linear DAEs the errors are known to be separated. To handle these problems, a defect correction generalizing the projection technique introduced in [2] for Hessenberg form index-2 DAEs is proposed.

**2. Simulation of electric circuits.** The simulation of electric circuits is of great interest today. The circuits we want to study here are assumed to be modelled by an RLC-network that can be devided into a dynamic network and a non-dynamic one, which are both connected by a b-gate. The non-dynamic network consists of linear resistors, nonlinear resistors, independent sources, and controlled sources. The

* supported by the Graduiertenkolleg "Geometrie und nichtlineare Analysis" der Humboldt–Universität zu Berlin

dynamic network contains linear and nonlinear capacitances and inductances. We speak of a nonlinear capacitance if there is a nonlinear differentiable mapping $Q_C = \psi(u_C)$ between the charge and voltage of the capacitance. Accordingly, we speak of a nonlinear inductance if there is a nonlinear differentiable mapping $\Phi_L = \varphi(I_L)$ between the flux and current of the inductance. Such networks can be modelled by differential algebraic equations (cf. [16]).

As usually, circuits consist of a large number of elements. The equations have to be generated automatically. Therefore, we want to study two modern modelling techniques making such an automatic generation possible, namely, the classical approach and the charge oriented approach of the modified nodal analysis (cf. [4], [5], [7], [8]).

The *classical modified nodal analysis* provides systems of the form

$$D(x)\dot{x} + f(x) = r(t), \tag{2.1}$$

where the vector of unknowns $x$ consists of

- the nodal potentials $u$ and
- the currents $I$ of the voltage-controlled elements.

The system contains the equations derived by Kirchhoff's nodal law for each node. Additionally, the characteristic equations of the voltage-controlled elements belong to the system. The equations of the current-controlled elements are set into the system directly.

The *charge oriented modified nodal analysis* leads to systems of the form

$$A\dot{q} + f(x) = r(t), \tag{2.2}$$
$$q - g(x) = 0. \tag{2.3}$$

Here, the vector of unknowns $(x, q)$ contains

- the nodal potentials $u$,
- the currents $I$ of the voltage-controlled elements,
- the charge $Q$ of the capacitors and
- the flux $\Phi$ of the inductors.

Equation (2.3) represents the characteristic equations for charge and flux.

Both modelling techniques are closely related. Denoting the derivative of the function $g$ with respect to $x$ by $g'(x)$, the relation

$$D(x) = Ag'(x) \tag{2.4}$$

is satisfied. The matrix $A$ is constant and its entries are numbers of the set $\{-1, 0, 1\}$. In general, this matrix is rectangular and not of full rank. The incidence matrix $A$ is proposed to be formulated properly such that

$$\operatorname{im} Ag'(x) \equiv \operatorname{im} A \tag{2.5}$$

becomes true. Then, the derivative-free equations in (2.1) are given by

$$(I - AA^+)(f(x) - r(t)) = 0,$$

while the derivative-free subsystem of (2.2)-(2.3) consists of

$$\begin{aligned} (I - AA^+)(f(x) - r(t)) &= 0, \\ q - g(x) &= 0. \end{aligned}$$

REMARK: If the network is modelled without inductances, (2.3) reads $g(x) = g(u)$. If the network contains neither a capacitance nor an inductance, i.e., if the circuit does not have dynamical elements, then equation (2.3) disappears completely, and the two modelling techniques lead to the same system $f(x) = r(t)$. Hence, we may exclude the latter case when studying the differences between both approaches.

For more clarity let us consider an example. Figure 1 displays a circuit simulating a NAND-gate (see [11]). It consists of two n-channel enhancement MOSFETs (ME), one n-channel depletion MOSFET (MD), and a load capacitor $C$ (cf. [18]).

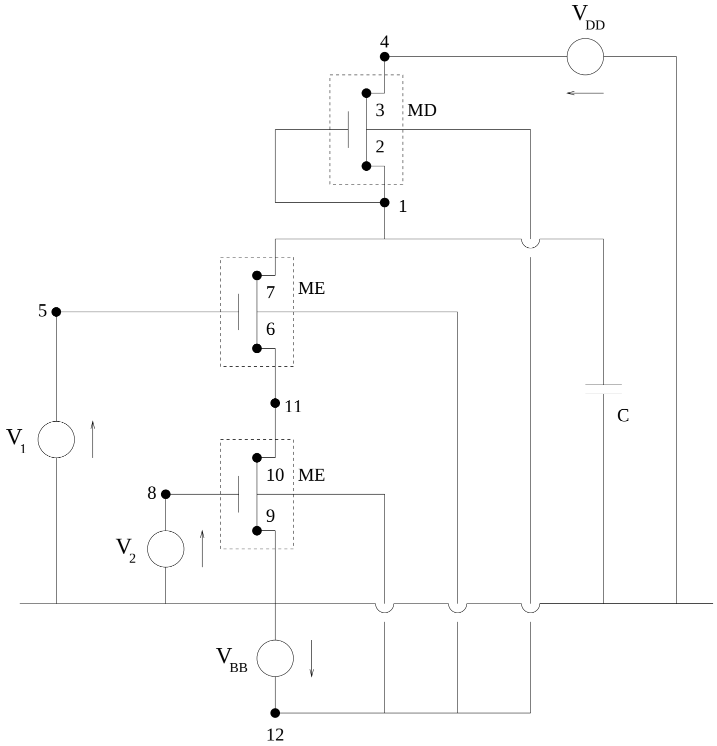


Fig.1 NAND-gate model

Digital MOS-circuits contain no other elements besides the MOSFETs as a rule. MOSFETs also take the function of controlled resistors. In our example, gate and source of the depletion transistor MD are connected, i.e., this MOSFET works as a controlled resistor here.

The drain voltage of MD is constant at $V_{DD} = 5V$. The bulk voltages are not at ground, $V_{BB} = -2.5V$. The source voltages of both MEs are at ground. The gate voltages are controlled by the voltage sources $V_1$ and $V_2$. The response is only LOW (FALSE) if both, the input signal $V_1$ and the input signal $V_2$ are HIGH (TRUE).

The circuit model for the MOSFETs MD and ME is given in Figure 2. This model is presented in [10] and leads to an index-2 system. It reflects the physical structure of the MOSFET well. However, note that the discussion on different MOSFET models is still going on, e.g. that on regularized versions leading to index 1 DAEs (e.g. [11]). The transistors MD and ME differ only in parameter values (see Table 1).

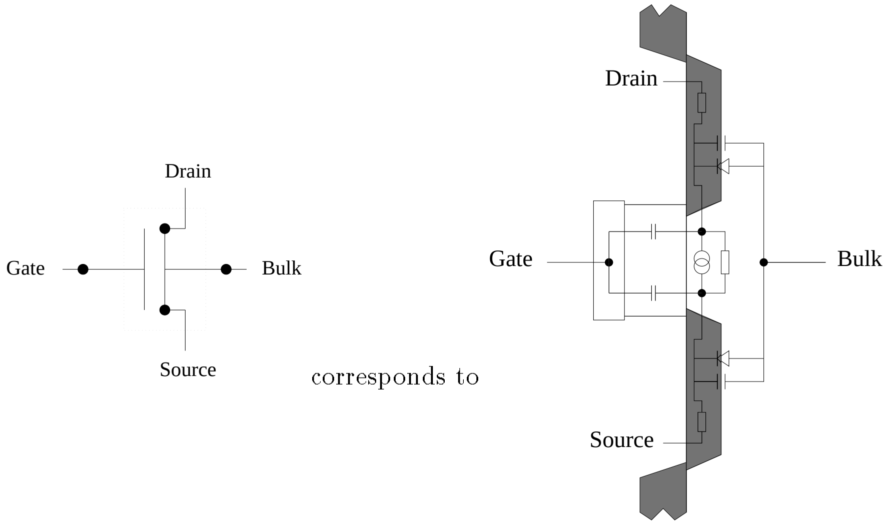


Fig. 2 MOSFET model

The current $i_{ds}$ flows from drain to source if and only if the controlling voltage $U_{gs}$ between gate and source is larger than a technology dependent threshold voltage $U_T$. The gate is isolated from the channel DS by a thin $SiO_2$-layer, i.e., the resistance $R_{sd}$ between source and drain is almost infinitely high ($\sim 10^{15}\Omega$).

Using the charge oriented modified nodal analysis, we obtain the following DAE system of dimension 29. Kirchhoff's nodal law applied to all nodals of the system leads to

$$
\begin{aligned}
\frac{u_1-u_2}{R_s} - \frac{u_7-u_1}{R_d} + \dot{Q} + \dot{Q}_{1gd} + \dot{Q}_{1gs} &= 0 \\
-\dot{Q}_{1gs} - \dot{Q}_{1bs} + \frac{u_2-u_1}{R_s} + \frac{u_2-u_3}{R_{sd}} + i_{bs}^{D}(u_{12}-u_2) & \\
+\, i_{ds}^{D}(u_3-u_2, u_1-u_2, u_{12}-u_2) &= 0 \\
-\dot{Q}_{1gd} - \dot{Q}_{1bd} + \frac{u_3-u_4}{R_d} - \frac{u_2-u_3}{R_{sd}} + i_{bd}^{D}(u_{12}-u_3) & \\
-\, i_{ds}^{D}(u_3-u_2, u_1-u_2, u_{12}-u_2) &= 0 \\
\frac{u_4-u_3}{R_d} + I_{DD} &= 0 \\
\dot{Q}_{2gd} + \dot{Q}_{2gs} + I_1 &= 0 \\
-\dot{Q}_{2gs} - \dot{Q}_{2bs} + \frac{u_6-u_{11}}{R_s} + \frac{u_6-u_7}{R_{sd}} + i_{bs}^{E}(u_{12}-u_6) & \\
+\, i_{ds}^{E}(u_7-u_6, u_5-u_6, u_{12}-u_6) &= 0 \\
-\dot{Q}_{2gd} - \dot{Q}_{2bd} + \frac{u_7-u_1}{R_d} - \frac{u_6-u_7}{R_{sd}} + i_{bd}^{E}(u_{12}-u_7) & \\
-\, i_{ds}^{E}(u_7-u_6, u_5-u_6, u_{12}-u_6) &= 0 \\
\dot{Q}_{3gd} + \dot{Q}_{3gs} + I_2 &= 0 \\
-\dot{Q}_{3gs} - \dot{Q}_{3bs} + \frac{u_9}{R_s} + \frac{u_9-u_{10}}{R_{sd}} + i_{bs}^{E}(u_{12}-u_9) & \\
+\, i_{ds}^{E}(u_{10}-u_9, u_8-u_9, u_{12}-u_9) &= 0
\end{aligned}
$$

$$
\begin{aligned}
-\dot{Q}_{3gd}-\dot{Q}_{3bd}+\frac{u_{10}-u_{11}}{R_d}-\frac{u_9-u_{10}}{R_{sd}}+i^E_{bd}(u_{12}-u_{10}) & \\
-i^E_{ds}(u_{10}-u_9,u_8-u_9,u_{12}-u_9) &= 0\\
\frac{u_{11}-u_6}{R_s}-\frac{u_{10}-u_{11}}{R_d} &= 0\\
\dot{Q}_{1bd}+\dot{Q}_{1bs}-i^D_{bs}(u_{12}-u_2)-i^D_{bd}(u_{12}-u_3) & \\
+\dot{Q}_{2bd}+\dot{Q}_{2bs}-i^E_{bs}(u_{12}-u_6)-i^E_{bd}(u_{12}-u_7) & \\
+\dot{Q}_{3bd}+\dot{Q}_{3bs}-i^E_{bs}(u_{12}-u_9)-i^E_{bd}(u_{12}-u_{10})+I_{BB} &= 0.
\end{aligned}
$$

The characteristic equations for the four voltage sources are given by

$$
\begin{aligned}
u_4-V_{DD} &= 0\\
u_{12}-V_{BB} &= 0\\
u_5-V_1 &= 0\\
u_8-V_2 &= 0.
\end{aligned}
$$

The characteristic equations for the capacitances of the system are described by

$$
\begin{aligned}
Q-C\,u_1 &= 0\\
Q_{1gd}-q_{gd}(u_1-u_3) &= 0\\
Q_{1gs}-q_{gs}(u_1-u_2) &= 0\\
Q_{1bd}-q_{bd}(u_{12}-u_3) &= 0\\
Q_{1bs}-q_{bs}(u_{12}-u_2) &= 0\\
Q_{2gd}-q_{gd}(u_5-u_7) &= 0\\
Q_{2gs}-q_{gs}(u_5-u_6) &= 0\\
Q_{2bd}-q_{bd}(u_{12}-u_7) &= 0\\
Q_{2bs}-q_{bs}(u_{12}-u_6) &= 0\\
Q_{3gd}-q_{gd}(u_8-u_{10}) &= 0\\
Q_{3gs}-q_{gs}(u_8-u_9) &= 0\\
Q_{3bd}-q_{bd}(u_{12}-u_{10}) &= 0\\
Q_{3bs}-q_{bs}(u_{12}-u_9) &= 0.
\end{aligned}
$$

The current through the diode between bulk and source as well as the current through the diode between bulk and drain is given by the function

$$
i_{bs}(U)=i_{bd}(U)=\begin{cases} -i_s\cdot\left(\exp(\frac{U}{U_T})-1\right) & \text{for } U\le 0\\ 0 & \text{for } U>0\end{cases}. \tag{2.6}
$$

The current through the controlled current source between drain and source is modelled by

$$
i_{ds}(U_{ds},U_{gs},U_{bs}) =
\begin{cases}
0 & \text{for } U_{gs}-U_{TE}\le 0\\
-\beta\cdot(1+\delta\cdot U_{ds})\cdot(U_{gs}-U_{TE}) & \text{for } 0<U_{gs}-U_{TE}\le U_{ds}\\
-\beta\cdot U_{ds}\cdot(1+\delta\cdot U_{ds})\cdot[2(U_{gs}-U_{TE})-U_{ds}] & \text{for } 0<U_{ds}<U_{gs}-U_{TE}
\end{cases}
$$

with $U_{TE}=U_{T0}+\gamma\cdot\left(\sqrt{\Phi-U_{bs}}-\sqrt{\Phi}\right)$. The technical parameters for the MOSFETs MD and ME are given in Table 1.

| | ME | MD |
|---|---|---|
| $i_s$ | $10^{-14}$A | $10^{-14}$A |
| $U_T$ | 25.85V | 25.85V |
| $U_{T0}$ | 0.8V | −2.43V |
| $\beta$ | $1.748 \cdot 10^{-3}$A/V$^2$ | $5.35 \cdot 10^{-4}$A/V$^2$ |
| $\gamma$ | $0.0\sqrt{\mathrm{V}}$ | $0.2\sqrt{\mathrm{V}}$ |
| $\delta$ | 0.02V$^{-1}$ | 0.02V$^{-1}$ |
| $\Phi$ | 1.01V | 1.28V |

Table 1: Technical parameters

The values for the resistances are chosen for all MOSFETs as

$$R_s = R_d = 4\Omega, \quad R_{sd} = 10^{15}\Omega.$$

The capacitance between gate and source as well as the capacitance between gate and drain are modelled as linear capacitors, i.e.,

$$q_{gs}(u) = q_{gd}(u) = C_1 \cdot u \text{ with } C_1 = 0.6 \cdot 10^{-13}F.$$

The capacitance between bulk and drain as well as the capacitance between bulk and source are modelled by nonlinear capacitances

$$q_{bd}(u) = q_{bs}(u) = \begin{cases} C_0 \cdot \Phi_B \cdot \left(1 - \sqrt{1 - \frac{u}{\Phi_B}}\right) & \text{for } u \le 0 \\ C_0 \cdot \left(1 + \frac{u}{4\Phi_B}\right) \cdot u & \text{for } u > 0 \end{cases}$$

with

$$C_0 = 0.24 \cdot 10^{-13}F \text{ and } \Phi_B = 0.87V.$$

Ordering the vector q as

$$q = (Q, Q_{1gd}, Q_{1gs}, Q_{1bd}, Q_{1bs}, Q_{2gd}, Q_{2gs}, Q_{2bd}, Q_{2bs}, Q_{3gd}, Q_{3gs}, Q_{3bd}, Q_{3bs})^T$$

we obtain for the incidence matrix

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The presented example shows that the charge oriented modelling technique leads to a system that is highly nonlinear and *not* of Hessenberg form. Further, the classical approach also leads to a system with these properties.

**3. Structure and index of electric circuits.** It is well-known that the numerical behaviour of integration methods for the solution of DAEs depends essentially on the *index* of the system. That's why the question whether both modelling techniques lead to the same index or not has been of great interest. This problem was already studied in [7] for some examples. In this section, we present some further results, in particular, for models whose capacitances are reciprocal one-port capacitances. In this case, each capacitance of the network has two uniquely determined nodals (including the node of the zero potential) enclosing this capacitance. That means, for each capacitance of the network the voltage through this capacitance may be expressed by the difference of the nodal potentials of these two uniquely determined nodals. For these models, the DAEs (2.1) and (2.2)-(2.3) have the following special structure

$$g'(x) = R(x)A^T \tag{3.1}$$

if the equations and variables are in proper order, and

$$R(x) = \begin{pmatrix} \psi_1'(x) & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \psi_2'(x) & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \psi_{n_C}'(x) & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \varphi_1'(x) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \varphi_2'(x) & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \varphi_{n_L}'(x) \end{pmatrix}$$

is symmetric and positive definite. The differentiable mappings $\psi_i$ and $\varphi_j$ describe the charge of the capacitance $C_i$ and the flux of the inductance $L_j$, respectively, for $i = 1, ..., n_C$ and $j = 1, ..., n_L$ (see page 2). Hence, the matrix $D(x)$ has the structure

$$D(x) = Ag'(x) = AR(x)A^T.$$

REMARK: The matrix $A$ may be written more precisely as

$$\mathrm{A} = \begin{array}{cc} \begin{pmatrix} M & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{pmatrix} & \begin{array}{c} n_u \\ n_L \\ n_s \end{array} \\ \begin{array}{ccc} n_C & n_L & n_s \end{array} & \end{array} \tag{3.2}$$

where $M$ is a matrix with the entries $-1$, 0, 1 only. It describes the occurrence of the capacitors in the network. $I$ represents the identity matrix. The dimension $n_s$ denotes the number of voltage controlled sources of the circuit. Let us remark that some dimensions (e.g. $n_L$) may be zero if the circuit contains not all kinds of elements. Then, obviously, some rows or columns disappear in the description (3.2).

LEMMA 3.1. *The model class described via (3.1) satisfies*

$$im\, A = im\, D(x) \quad and \quad ker\, D(x) = ker\, g'(x) = ker\, A^T.$$

*In both systems (2.1) and (2.2)-(2.3), the leading coefficients* $D(x)$ *resp.* $\begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}$ *have constant nullspaces.*

*Proof.* Since $g'(x) = R(x)A^T$ is valid for a symmetric positive definite matrix $R(x)$, we find a symmetric regular matrix $R_s(x)$ such that

$$R(x) = R_s(x)R_s(x)$$

is satisfied. Hence,

$$\operatorname{rank} AR(x)A^T = \operatorname{rank} AR_s(x)(AR_s(x))^T = \operatorname{rank} AR_s(x) = \operatorname{rank} A.$$

This implies $\operatorname{im} D(x) = \operatorname{im} AR(x)A^T = \operatorname{im} A$. Secondly,

$$\operatorname{rank} AR(x)A^T = \operatorname{rank} AR_s(x)(AR_s(x))^T = \operatorname{rank}(AR_s(x))^T = \operatorname{rank} R(x)A^T.$$

Now, the relation $\ker D(x) = \ker AR(x)A^T = \ker g'(x)$ holds. Taking into account that $R(x)$ is a nonsingular matrix we obtain the relation

$$\ker g'(x) = \ker R(x)A^T = \ker A^T.$$

Hence, it follows that $\ker D(x) = \ker A^T$ is constant. □

Next, for investigating the tractability index (cf. [13]) of the two systems (2.1) and (2.2)-(2.3), we introduce the characteristic linear subspaces

$$\begin{aligned} N(x) &:= \ker D(x) \subseteq \mathbb{R}^m, \\ S(x) &:= \{z : f'(x)z \in \operatorname{im} D(x) = \operatorname{im} A\} \subseteq \mathbb{R}^m, \end{aligned}$$

which are related to (2.1), and further

$$\begin{aligned} \tilde{N} &:= \{\left(\begin{smallmatrix} \gamma \\ z \end{smallmatrix}\right) : A\gamma = 0\} = \ker A \times \mathbb{R}^m \subseteq \mathbb{R}^{m+n}, \\ \tilde{S}(x) &:= \{\left(\begin{smallmatrix} \gamma \\ z \end{smallmatrix}\right) : f'(x)z \in \operatorname{im} A,\ \gamma = g'(x)z\}. \end{aligned}$$

In our context (cf. (2.5)), the two leading nullspaces $N(x)$ and $\tilde{N}$ have constant dimension, that is, $\dim N(x) = m - r$, $\dim \tilde{N} = m + n - r$, where $r := \operatorname{rank} A$.

Now, we are prepared to apply the well-known criteria for the index-1 tractability (transferability) of our DAEs ([9], [13]). More precisely, (2.1) has index 1 if

$$N(x) \cap S(x) = \{0\}$$

holds true for all $x$ and, similarly, (2.2)-(2.3) is an index 1 system if

$$\tilde{N} \cap \tilde{S}(x) = \{0\}.$$

Do both model classes lead to the same index? Compute

$$\begin{aligned} \tilde{N} \cap \tilde{S}(x) &= \{\left(\begin{smallmatrix} \gamma \\ z \end{smallmatrix}\right) : A\gamma = 0,\ f'(x)z \in \operatorname{im} A,\ \gamma = g'(x)z\} \\ &= \{\left(\begin{smallmatrix} \gamma \\ z \end{smallmatrix}\right) : \gamma = g'(x)z,\ Ag'(x)z = 0,\ f'(x)z \in \operatorname{im} Ag'(x)\} \\ &= \{\left(\begin{smallmatrix} \gamma \\ z \end{smallmatrix}\right) : \gamma = g'(x)z,\ z \in N(x) \cap S(x)\}, \end{aligned}$$

hence, $\dim \tilde{N}\cap\tilde{S}(x) = \dim N(x)\cap S(x)$. Obviously, both systems (2.1) and (2.2)-(2.3) are index-1 tractable simultaneously.

However, the classical MNA system (2.1) may have a leading nullspace $N(x)$ rotating with $x$ while the charge oriented version (2.2)-(2.3) leads always to the constant nullspace $\tilde{N}$. Recall that an index-1 tractable DAE having a leading nullspace varying with the solution behaves analytically and numerically like an index-2 tractable DAE with constant nullspace. It has the perturbation index 2 (cf. [14]) then.

Note that the so-called general capacitance interpretation may lead, in fact, to a non-symmetric matrix $D(x)$ the nullspace of which varies with $x$ ([6]). Hence, from the point of view of DAE theory, the charge/flux-oriented formulation has a great advantage. Due to the constant leading nullspace, the tractability index 1 of (2.2)-(2.3) implies the perturbation index 1 whereas the perturbation index of (2.1) is 2.

For a detailed analysis of quasilinear index-1 DAEs whose leading nullspace varies with $x$ we refer to [1], [14]. Fortunately, if the nullspace does not rotate to fast, the instabilities in numerical integrations caused by the higher perturbation index behave very weakly.

In the present paper we are mainly interested in equations having tractability index 2. We specify criteria resp. results for both (2.1) and (2.2)-(2.3) in more detail. However, index-2 tractability has been defined and investigated for the case of constant leading nullspace only, yet. This is why we also assume

$$\ker D(x) = N(x) = N \tag{3.3}$$

to be constant in the following. Note that this assumption is trivially satisfied in the symmetric case described by (3.1).

Now, for investigating the tractability index 2 (cf. [13]) we choose constant projectors $Q$ onto $\ker D(x)$ and $Q_A$ onto $\ker A$, respectively. Furthermore, we define $P := I - Q$, $P_A := I - Q_A$ and introduce the linear subspaces

$$\begin{aligned} N_1(x) &:= \ker [D(x) + f'(x)Q] \subseteq \mathbb{R}^m, \\ S_1(x) &:= \{z : \ f'(x)Pz \in \operatorname{im} D(x) = \operatorname{im} [D(x) + f'(x)Q]\} \subseteq \mathbb{R}^m, \end{aligned}$$

which are related to (2.1) as well as

$$\begin{aligned} \tilde{N}_1(x) &:= \{\left(\begin{smallmatrix}\gamma\\ z\end{smallmatrix}\right) : \ A\gamma + f'(x)z = 0,\ Q_A\gamma = g'(x)z\} \subseteq \mathbb{R}^{m+n}, \\ \tilde{S}_1(x) &:= \{\left(\begin{smallmatrix}\gamma\\ z\end{smallmatrix}\right) : \ \exists\,\alpha,\beta : \ 0 = A\alpha + f'(x)\beta,\ P_A\gamma = Q_A\alpha - g'(x)\beta\}, \end{aligned}$$

which are related to (2.2)-(2.3). Then, (2.1) is index-2 tractable if

$$N_1(x) \cap S_1(x) = \{0\} \quad \text{and} \quad \dim [N(x) \cap S(x)] = \text{const} > 0$$

are fulfilled for all $x$. Correspondingly, the system (2.2)-(2.3) is index-2 tractable if

$$\tilde{N}_1(x) \cap \tilde{S}_1(x) = \{0\} \quad \text{and} \quad \dim [\tilde{N} \cap \tilde{S}(x)] = \text{const} > 0.$$

Recall our notion of index-2 tractability to be a straightforward generalization of the corresponding definition for the linear case, which, in its turn, represents a generalization of the Kronecker index. On the other hand, nonlinear index-2 Hessenberg systems are known to be index-2 tractable, too.

THEOREM 3.2. *The model class described by (2.5) and (3.3) satisfies the following assertion.*
*The system (2.1) is index-2 tractable if and only if the system (2.2)-(2.3) is so.*

REMARKS:

1. Both modelling techniques lead to the same index for the lower index case.
2. Since the nullspaces of the leading coefficients are constant but do not rotate, we may expect integration methods to work well (cf. [13], [20]).
3. The network equation system of the NAND-gate example above is index-2 tractable (see [21]). Moreover, it has the differential index 2 (see [7], [10]) and the perturbation index 2 (apply Theorem 4.5).

*Proof.* We have already seen above that $\tilde{N} \cap \tilde{S}(x)$ has the same dimension as $N(x) \cap S(x)$. Therefore, it is sufficient to prove the assertion

$$N_1(x) \cap S_1(x) = \{0\} \quad \leftrightarrow \quad \tilde{N}_1(x) \cap \tilde{S}_1(x) = \{0\}.$$

($\rightarrow$) For any $\binom{\gamma}{\tilde{z}} \in \tilde{N}_1(x) \cap \tilde{S}_1(x)$, there exist $\alpha$, $\beta$ such that

$$P_A\gamma = Q_A\alpha - g'(x)\beta \tag{3.4}$$

$$0 = A\alpha + f'(x)\beta \tag{3.5}$$

are satisfied. Because of $\binom{\gamma}{\tilde{z}} \in \tilde{N}_1(x)$, we may conclude

$$\tilde{z} \in \ker Ag'(x) = \operatorname{im} Q$$

if we regard $Q_A\gamma = g'(x)\tilde{z}$. We introduce $z := \tilde{z} - P\beta$. Using (3.4), we obtain

$$Ag'(x)z = Ag'(x)\tilde{z} - Ag'(x)\beta = -Ag'(x)\beta = A\gamma.$$

Since $\binom{\gamma}{\tilde{z}} \in \tilde{N}_1(x)$, the relation

$$-A\gamma = f'(x)\tilde{z} = f'(x)Q\tilde{z} = f'(x)Qz$$

is fulfilled. The latter two equations lead to

$$z \in N_1(x). \tag{3.6}$$

Further, we obtain

$$\begin{aligned} f'(x)Pz &= -f'(x)P\beta = -f'(x)\beta + f'(x)Q\beta \\ &= A\alpha + f'(x)Q\beta \qquad \text{(see (3.5))} \\ &= Ag'(x)\alpha_1 + f'(x)Q\beta \quad \text{(for a certain } \alpha_1\text{, since } \operatorname{im} A = \operatorname{im} Ag'\text{)} \\ &= (Ag'(x) + f'(x)Q)(P\alpha_1 + Q\beta), \end{aligned}$$

that means,

$$z \in S_1(x). \tag{3.7}$$

Now, (3.6) and (3.7) imply $z = 0$. Because of $\tilde{z} = Q\tilde{z}$, the relation $\tilde{z} = Qz = 0$ is valid. Further, from (3.4) we conclude that

$$A\gamma = -Ag'(x)\beta = Ag'(x)Pz = 0$$

is satisfied. Finally, $\gamma = Q_A\gamma = g'(x)\tilde{z} = 0$, i.e.,

$$\tilde{N}_1(x) \cap \tilde{S}_1(x) = \{0\}.$$

($\leftarrow$) For any $z \in N_1(x) \cap S_1(x)$, we find an $\alpha_1$ such that

$$f'(x)Pz = Ag'(x)\alpha_1 + f'(x)Q\alpha_1. \tag{3.8}$$

We consider

$$\begin{array}{ll} \gamma & := P_A g'(x)z + g'(x)\tilde{z}, \quad \tilde{z} := Qz, \\ \alpha & := P_A g'(x)\alpha_1 + Q_A g'(x)\beta, \quad \beta := Q\alpha_1 - Pz. \end{array}$$

Then,

$$\begin{aligned} A\gamma + f'(x)\tilde{z} &= Ag'(x)z + Ag'(x)\tilde{z} + f'(x)Qz \\ &= (Ag'(x) + f'(x)Q)z = 0 \\ Q_A\gamma &= Q_A g'(x)\tilde{z} = g'(x)\tilde{z} \quad (\text{since } P_A g'(x)Q = 0). \end{aligned}$$

Hence,

$$\begin{pmatrix} \gamma \\ \tilde{z} \end{pmatrix} \in \tilde{N}_1(x) \tag{3.9}$$

is satisfied. Further,

$$\begin{aligned} 0 &= Ag'(x)\alpha_1 + f'(x)Q\alpha_1 - f'(x)Pz = A\alpha + f'(x)\beta \quad (\text{see } (3.8)) \\ P_A\gamma &= P_A g'(x)z = P_A g'(x)Pz = -P_A g'(x)P\beta = -P_A g'(x)\beta \\ &= Q_A g'(x)\beta - g'(x)\beta = Q_A\alpha - g'(x)\beta, \end{aligned}$$

i.e., $\begin{pmatrix} \gamma \\ \tilde{z} \end{pmatrix} \in \tilde{S}_1(x)$. Together with (3.9)), this leads to

$$\begin{pmatrix} \gamma \\ \tilde{z} \end{pmatrix} \in \tilde{N}_1(x) \cap \tilde{S}_1(x),$$

i.e., $\gamma = \tilde{z} = 0$. Now, we know

$$P_A g'(x)z = 0, \quad Qz = 0.$$

The first relation implies $z \in \ker Ag'(x)$, i.e., $z \in \operatorname{im} Q$. Together with the second relation, we conclude that $z = 0$, i.e.,

$$N_1(x) \cap S_1(x) = 0.$$

$\square$

**4. Linearizations and solvability.** In this section we give deeper insight into the analytical background of (2.1) and (2.2)-(2.3). The functions $f$ and $g$ involved in (2.1) and (2.2)-2.3 are supposed to be continuously differentiable on their definition domain $\mathcal{D} \subseteq \mathbb{R}^m$.
Due to (3.3), equation (2.1) can be rewritten more precisely as

$$D(x(t))\frac{d}{dt}(Px(t)) + f(x(t)) - r(t) = 0, \tag{4.1}$$

whereby $P \in L(\mathbb{R}^m)$ denotes any constant projector matrix projecting along the constant nullspace $N = \ker D(x)$. This reformulation (4.1) provides information on what kind of functions we should accept to be solutions of the DAE (2.1) in fact. Namely, such a solution has to be a continuous function with a continuously differentiable $P$-component. However, the other component should not be expected to belong to $C^1$ in general.

Analogously, the system (2.2)-(2.3) means more precisely

$$\begin{aligned} A\frac{d}{dt}(P_A q(t)) + f(x(t)) &= r(t), \\ q(t) - g(x(t)) &= 0, \end{aligned}$$

whereby $P_A \in L(\mathbb{R}^m)$ denotes any constant projector matrix projecting along the $\ker A$. Hence, the function spaces

$$C^1_N := \{x \in C(\mathcal{J}, \mathbb{R}^m) : Px \in C^1(\mathcal{J}, \mathbb{R}^m)\},$$

$$C^1_{\tilde{N}} := \{\tilde{x} = (x, q) \in C(\mathcal{J}, \mathbb{R}^{m+n}) : P_A q \in C^1(\mathcal{J}, \mathbb{R}^n)\}$$

result to be natural ones which the solutions of (2.1) resp. (2.2)-(2.3) should belong to. $\mathcal{J} \subseteq \mathbb{R}$ denotes the given interval.

Remark: The projector $P_A$ is easy to compute because of the very special structure of $A$.

Our first assertion answers the question on the equivalence of the systems (2.1) and (2.2)-(2.3).

Theorem 4.1. *$(x, q) \in C^1_{\tilde{N}}$ is a solution of (2.2)-(2.3) if and only if $x \in C^1_N$ solves (2.1) and $q(t) = g(x(t), t)$, $t \in \mathcal{J}$.*

*Proof.* Denote again $Q := I - P$. Clearly, $Q$ projects onto $N$. The special structure of $g$ leads to the relation

$$Ag(x) - Ag(Px) = \int_0^1 Ag'(sx + (1-s)Px)Q\, ds = 0,$$

i.e., $Ag(x) = Ag(Px)$, $x \in \mathcal{D}$.

Now, given a solution $x \in C^1_N$ of (2.1) and $q(t) = g(x(t))$, $t \in \mathcal{J}$. Because of $Ag(x(t)) \equiv Ag(Px(t))$ and $Px \in C^1$, the function $P_A q$ is continuously differentiable, too. In particular, we have $\frac{d}{dt}(P_A q(t)) = P_A g'(x(t)) \frac{d}{dt}(Px(t))$. Thus, the pair $(x, q)$ belongs to $C^1_{\tilde{N}}$ and satisfies (2.2)-(2.3).
On the contrary, for a given solution $(x, q) \in C^1_{\tilde{N}}$ of (2.2)-(2.3) we have $x \in C$, $q \in C$, $P_A q \in C^1$. In more detail, the relation $Ag(x(t)) = Ag(Px(t))$ shows

$$P_A q(t) = P_A g(Px(t)).$$

The matrix function $Ag'(x)$ has the constant nullspace $N$ and the constant range $\operatorname{im} A$. Applying the Implicit Function Theorem we find the function $Px$ to be as smooth as $P_A q$. Consequently, now have $Px \in C^1$, $x \in C^1_N$. Obviously, the DAE (2.1) is satisfied. $\square$

COROLLARY 4.2. *If* $(x, q) \in C^1_{\tilde{N}}$ *solves (2.2)-(2.3), then we always have* $Px \in C^1$.

Denote by $Q_*(x) \in L(\mathbb{R}^m)$ the orthoprojector onto $S(x)$, $x \in \mathcal{D}$. Recall the possible representation

$$Q_*(x) = I - [(I - AA^+)f'(x)]^+(I - AA^+)f'(x).$$

THEOREM 4.3. *Let the subspaces* $S(x) \subset \mathbb{R}^m$ *and* $S(x) \cap N \subset \mathbb{R}^m$, $x \in \mathcal{D}$, *have constant dimensions* $r$ *and* $\mu$, *respectively,* $r := \text{rank } A$.
*Then, the system (2.2)-(2.3) is index-2 tractable if and only if, for* $x \in \mathcal{D}$,

$$z \in N \cap S(x),\ f'(x)z \in im\, Ag'(x)Q_*(x) \qquad imply \qquad z = 0. \tag{4.2}$$

*Proof.* Reformulate (2.2)-(2.3) in standard DAE form

$$\tilde{A}\dot{\tilde{x}}(t) + \tilde{g}(\tilde{x}(t)) - \tilde{r}(t) = 0, \tag{4.3}$$

with a quadratic leading coefficient matrix $\tilde{A}$,

$$\tilde{A} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{x} = \begin{pmatrix} q \\ x \end{pmatrix}, \quad \tilde{g}(\tilde{x}) = \begin{pmatrix} f(x) \\ q - g(x) \end{pmatrix}.$$

Introduce the projector $\tilde{Q} := \begin{pmatrix} Q_A & 0 \\ 0 & I \end{pmatrix} \in L(\mathbb{R}^{m+n})$ onto the nullspace $\tilde{N} = \ker \tilde{A}$.

Next, define for $x \in \mathcal{D}$

$$\tilde{A}_1(x) = \tilde{A} + \tilde{g}'_{\tilde{x}}(\tilde{x})\tilde{Q} = \begin{pmatrix} A & f'(x) \\ Q_A & -g'(x) \end{pmatrix},$$

whose nullspace reads

$$\tilde{N}_1(x) = \ker \tilde{A}_1(x) = \{\left(\begin{smallmatrix} \gamma \\ z \end{smallmatrix}\right) \in \mathbb{R}^{m+n} : z \in N \cap S(x),\ \gamma = -A^+ f'(x)z\}.$$

Hence, this nullspace has also constant dimension $\dim \tilde{A}_1(x) = \dim S(x) \cap N = \mu$. By the definitions (e.g. [14]), (4.3) (resp. (2.2)-(2.3)) is index-2 tractable if $\tilde{A}_1(x)$ has constant rank $< m + n$ and $\ker \tilde{A}_1(x) \cap \tilde{S}_1(x) = \{0\}$,

$$\begin{aligned} \tilde{S}_1(x) &= \{\tilde{z} = \left(\begin{smallmatrix} \gamma \\ z \end{smallmatrix}\right) \in \mathbb{R}^{m+n} : \tilde{g}'_{\tilde{x}}(\tilde{x})\tilde{P}\tilde{z} \in \text{im } \tilde{A}_1(x)\} \\ &= \{\left(\begin{smallmatrix} \gamma \\ z \end{smallmatrix}\right) \in \mathbb{R}^{m+n} : A\gamma \in \text{im}\, Ag'(x)Q_*(x)\}. \end{aligned}$$

Note that $\tilde{N}_1(x)$ as well as $\tilde{S}_1$ are exactly the subspaces introduced in Section 3, however, in different representation.

In particular, for $\left(\begin{smallmatrix} \gamma \\ z \end{smallmatrix}\right) \in \ker \tilde{A}_1(x) \cap \tilde{S}_1(x)$ it holds that $z \in S(x)$, hence

$$A\gamma = -AA^+ f'(x)z = -f'(x)z.$$

Now the assertion follows immediately. $\square$

Let us turn to a linearization of (2.2)-(2.3) taken along a fixed function $\tilde{x}_* = (q_*, x_*) \in C^1_{\tilde{N}}$ whose trajectory remains in $\mathcal{D} \times \mathbb{R}^n$. Consider the linearized DAE

$$A\dot{p}(t) + f'(x_*(t))z(t) = s_1(t), \tag{4.4}$$

$$p(t) - g'(x_*(t))z(t) = s_2(t), \tag{4.5}$$

which is also index-2 tractable if (2.2)-(2.3) is so (cf. [14]). Unfortunately, the reverse is not true in general. (4.4)-(4.5) may have index 2 whereas (2.2)-(2.3) is rather a singular index-1 problem. This kind of singularities needs some special effort even in view of numerical computations (e.g. [17], [22]). In the present paper we avoid these situations by supposing additional structural conditions ensuring to have index-2 tractability in a neighbourhood of the trajectory of $\tilde{x}_*$, too.

For more clarity, we restrict ourselves to specifying the structural condition discussed in [15] and [14]. Further conditions will be considered in [21]. The next Lemma follows immediately from Lemma 4.1 in [14].

Lemma 4.4. *Let (4.4)-(4.5) be index-2 tractable and let the matrix*

$$\mathcal{M}(y) := \begin{pmatrix} 0 & (I - AA^+)f'(y) \\ Q_A & -g'(y) \end{pmatrix}, \qquad y \in \mathcal{D} \tag{4.6}$$

*have constant range.*
*Then (4.4)-(4.5) is index-2 tractable at least in a neighbourhood of the trajectory of $x_*$, i.e., the system (2.2)-(2.3) is index-2 tractable in this neighbourhood.*

In the very special case of (2.2)-(2.3) being linear, that is, $f'(y) \equiv F$, $g'(y) \equiv G$, the matrix $\mathcal{M}(y)$ has constant range, trivially. However, even in this case (2.2)-(2.3) is not in Hessenberg form.

Theorem 4.5. *Given a solution $\tilde{x}_* = (q_*, x_*) \in C^1_{\tilde{N}}$ of (2.2)-(2.3), $\mathcal{J}$ a compact interval, $t_0 \in \mathcal{J}$. Let the DAE linearized along $\tilde{x}_*$ be index-2 tractable, and let the matrix $\mathcal{M}(y)$ have a constant range. Moreover, let $(I - AA^+)f$ and $g$ be twice continuously differentiable but $(I - AA^+)r \in C^2(\mathcal{J}, \mathbb{R}^m)$.*

*(i) Then, the perturbed initial value problems*

$$\begin{aligned} A\dot{q}(t) + f(x(t)) - r(t) &= \rho(t), \\ q(t) - g(x(t)) &= 0, \\ \Pi(t_0)(q(t_0) - q^0) &= 0 \end{aligned} \tag{4.7}$$

*are uniquely solvable on $C^1_{\tilde{N}}(\mathcal{J}, \mathbb{R}^{m+n})$ supposed $|\Pi(t_0)(q^0 - q_*(t_0)|$ as well as $\|\rho\|_\infty + \|\frac{d}{dt}(\Omega\rho)\|_\infty$ are sufficiently small, $\rho \in C(\mathcal{J}, \mathbb{R}^m)$, $\Omega\rho \in C^1(\mathcal{J}, \mathbb{R}^m)$, $q_0 \in \mathbb{R}^n$. Thus, $\Pi(t_0)$ and $\Omega(t)$ are certain matrices described below (cf. (4.8), (4.10)).*

*(ii) For the solution of (i) the inequality*

$$\begin{aligned} &\|x - x_*\|_\infty + \|q - q_*\|_\infty + \|\frac{d}{dt}(P_A q) - \frac{d}{dt}(P_A q_*)\|_\infty \\ &\quad \le K\{\|\rho\|_\infty + \|\tfrac{d}{dt}(\Omega\rho)\|_\infty + |\Pi(t_0)(q(t_0) - q_*(t_0))|\} \end{aligned}$$

*is given with a constant $K > 0$.*

*Proof.* The assertion follows from [14], Theorem 4.4 by providing the right projectors used therein for our system (2.2)-(2.3). Choosing a continuous matrix function $R_*(x)$ (e.g. $[Ag'(x)]^+A$) satisfying

$$Ag'(x)R_*(x) = A \quad \text{and} \quad R_*(x)P_A = R_*(x),$$

and regarding the relations

$$PR_*g' = P, \qquad Q_1Q_* = 0,$$

the canonical projector $\tilde{Q}_1(x)$ onto $\tilde{N}_1$ along $\tilde{S}_1$ has the form

$$\tilde{Q}_1(x) = \begin{pmatrix} P_A g'(x)Q_1(x)R_*(x) & 0 \\ QQ_1(x)R_*(x) & 0 \end{pmatrix},$$

where $Q_1$ is the canonical projector $Q_1$ onto $N_1$ along $S_1$. We refer to [21] for technical computations. Now, we have

$$\tilde{P}\tilde{P}_1(x) = \begin{pmatrix} P_A - P_A g'(x)Q_1(x)R_*(x) & 0 \\ 0 & 0 \end{pmatrix},$$

but, in particular

$$\tilde{P}\tilde{P}_1(x_*(t_0)) = \begin{pmatrix} P_A - P_A g'(x_*(t_0))Q_1(x_*(t_0))R_*(x_*(t_0)) & 0 \\ 0 & 0 \end{pmatrix}.$$

It follows that

$$\Pi(t_0) = P_A - P_A g'(x_*(t_0))Q_1(x_*(t_0))R_*(x_*(t_0)) \tag{4.8}$$

should be chosen to state the initial condition. $\Pi(t_0)$ may be shown to be a projector again. Furthermore, to apply Theorem 4.4 mentioned above we use the representation

$$\tilde{Q}_1(x)\tilde{G}_2^{-1}(x)\begin{pmatrix} \rho \\ 0 \end{pmatrix} = \begin{pmatrix} P_A g'(x)Q_1(x)H^{-1}(x)\rho \\ QQ_1(x)H^{-1}(x)\rho \end{pmatrix}, \tag{4.9}$$

where

$$H(x) := D(x) + f'(x)[Q + PQ_1(x)]$$

is regular since the system (2.1) is index-2 tractable. Hence, with (4.8) and

$$\Omega(t) := Q_1(x_*(t))H^{-1}(x_*(t)) \tag{4.10}$$

our assertion follows immediately from [14], Theorem 4.4. □

REMARKS:

1. The inequality *(ii)* shows the perturbation index of (2.2)-(2.3) also to be 2 (cf. [12]).

2. The projector $\Pi(t_0)$ gives an idea of which of the variables involved in (2.2)-(2.3) are actually the state variables. Since in the index-2 case, we have an additional hidden constraint besides the obvious constraint

$$\begin{aligned} (I - AA^+)\{f(x(t)) - r(t) - \rho(t)\} &= 0, \\ q(t) - g(x(t)) &= 0, \end{aligned}$$

we cannot further expect $P_A q$ to be the state variable, but only a certain part of it.

3. In numerical computations the components

$$\tilde{P}\tilde{Q}_1 = \begin{pmatrix} P_A g' Q_1 R_* & 0 \\ 0 & 0 \end{pmatrix}$$

play an important role. These are the components that are subjected to an inherent differentiation causing numerical difficulties (cf. Section 5).

The solution $\tilde{x}(t, p^0)$ of the initial value problem (2.2)-(2.3), (4.7) depends continuously on $p^0$, but the partial derivative

$$\tilde{Z}(t) := \frac{\partial \tilde{x}}{\partial p^0}(t, p^0) = \begin{pmatrix} \Gamma(t) \\ Z(t) \end{pmatrix}$$

satisfies the first variation system

$$\begin{aligned} A\Gamma'(t) + f'(x(t,p^0))Z(t) &= 0, \\ \Gamma(t) - g'(x(t,p^0))Z(t) &= 0, \\ \Pi(t_0)(\Gamma(t_0) - I) &= 0. \end{aligned}$$

This makes clear that linearization works well in this situation, too. However, we should keep in mind that the sensitivity matrix $\tilde{Z}(t)$ does not have full rank, but

$$\begin{aligned} \ker \tilde{Z}(t) &\equiv \ker \Pi(t_0), \\ \dim \ker \Pi(t_0) &= r - \mu. \end{aligned}$$

In a similar way we may treat also the equations (2.2)-(2.3) which depend on additional parameters.

**5. BDF.** One of the most frequently used methods is the BDF, which we want to investigate in more detail for the special systems of the circuit simulation (2.2)-(2.3). Supposed the system (2.2)-(2.3) has index 1, the BDF is known to work well. Good experience on treating index 2 Hessenberg form DAEs is reported e.g. in [3]. However, what about the BDF applied to those nonlinear index 2 DAEs (2.2)-(2.3) that are not in Hessenberg form?

In the following we specify the more general considerations given in [15] and [20] to the circuit simulation systems (2.2)-(2.3) that are of interest here.

Let $(q_*, x_*)$ be a solution of the system (2.2)-(2.3) and let the DAE (2.2)-(2.3) be index-2 tractable locally around $(q_*, x_*)$. Further, let $\pi$ be a partition of the closed interval $[t_0, T]$ with the following properties

$$\begin{aligned} &\pi : t_0 < t_1 < \ldots < t_N = T \\ &0 < h_{min} \le t_\ell - t_{\ell-1} \le h_{max}, \quad \ell \ge 1, \\ &\kappa_1 \le \tfrac{h_{\ell-1}}{h_\ell} \le \kappa_2, \quad \ell \ge 2, \end{aligned} \tag{5.1}$$

where $\kappa_1$ and $\kappa_2$ are suitable constants and the variable stepsize and variable order BDF is stable for explicit ODE's.

The variable order, variable stepsize BDF for DAEs of the form (2.2)-(2.3) reads

$$A\frac{1}{h_\ell}\sum_{i=0}^{k} \alpha_{\ell i} q_{\ell - i} + f(x_\ell) - r(t_\ell) = \delta_\ell, \tag{5.2}$$

$$q_\ell - g(x_\ell) = 0. \tag{5.3}$$

Here, $\delta_\ell$ represents the perturbations in the $\ell$-th step caused by the rounding errors and the defects arising when solving the nonlinear equations numerically (e.g. by the Newton method). Applying Theorem 3.1 of [20] and regarding relation (4.9) we obtain the following result.

THEOREM 5.1. *Let the assumptions of Theorem 4.5 be fulfilled. Supposed there is a constant $C > 0$ such that the starting values satisfy the relation*

$$\|q_\ell - q_*(t_\ell)\| \le C h_\ell, \quad \ell < k,$$

*the following statements are true:*

*(i) There are constants $\epsilon > 0$ and $r > 0$ so that for all partitions (5.1) with sufficiently small stepsizes the BDF with*

$$\|\delta_\ell\| \le \epsilon, \quad \ell \ge k \quad \text{and} \quad \|Q_{1\ell} H_\ell^{-1} \delta_\ell\| \le h_\ell\, \epsilon, \quad \ell \ge 0$$

*is feasible in a neighbourhood of the trajectory $(q_*, x_*)$ with a constant radius $r$.*

*(ii) Supposed there is a constant $C_1 > 0$ with*

$$\|\delta_\ell\| \le C_1 h_\ell,\ \ell \ge k, \quad \|Q_{1\ell} H_\ell^{-1} \delta_\ell\| \le C_1 h_\ell^2,\ \ell \ge 0,$$

*we find a constant $C_2 > 0$ such that the following error estimation holds:*

$$\max_{\ell \ge k} \left\| \begin{pmatrix} x_\ell - x_*(t_\ell) \\ q_\ell - q_*(t_\ell) \end{pmatrix} \right\| \le C_2 \Big[ \max_{\ell < k} \|q_\ell - q_*(t_\ell)\| + \max_{\ell \ge k} \|\tau_\ell\|$$

$$+ \max_{\ell \ge k} \|\delta_\ell\| + \max_{\ell \ge 0} \tfrac{1}{h_\ell} \|Q_{1\ell} H_\ell^{-1} \delta_\ell\| \Big],$$

*where $\tau_\ell$ represents the local discretization error.*

Recall that we have

$$\tau_\ell = A \left\{ \frac{1}{h_\ell} \sum_{i=0}^{k} \alpha_{\ell i} q_*(t_{\ell - i}) - q_*'(t_\ell) \right\}.$$

Feasibility means that the nonlinear equations to be solved per integration step are locally uniquely solvable, and the Newton method applies.

Let us return to the above model of the NAND-gate. We have tested the variable order and variable stepsize BDF with the input voltages shown in Figure 3.

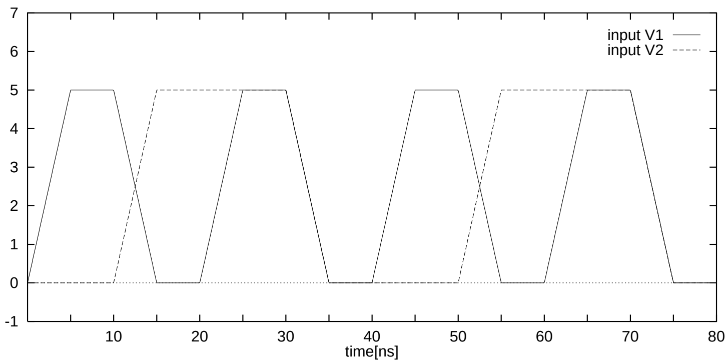


Fig. 3: Input signals $V_1$ and $V_2$

The simulation results reflect the real output of the NAND-gate. The voltage $u_1$ at node 1 is low if and only if the input voltages $V_1$ and $V_2$ are high. Figure 4 shows the numerical results. The regions $[10ns, 15ns]$ and $[50ns, 55ns]$ are critical. Both signals, $V_1$ and $V_2$ are relatively high around the time points 12.5ns and 52.5ns.

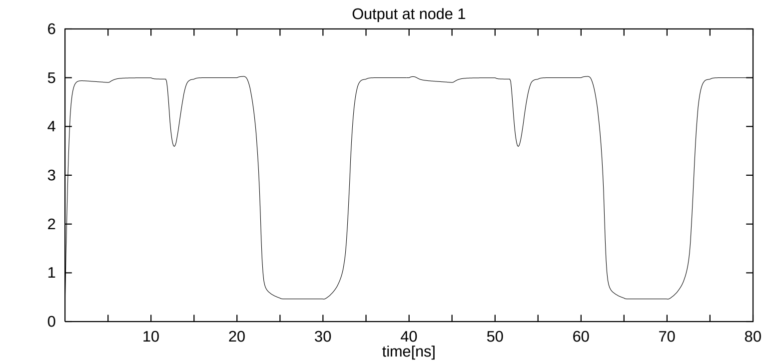


Fig. 4: Response at node 1

Figure 5 shows the result for $I_1$ and Figure 6 shows the result for $I_2$.

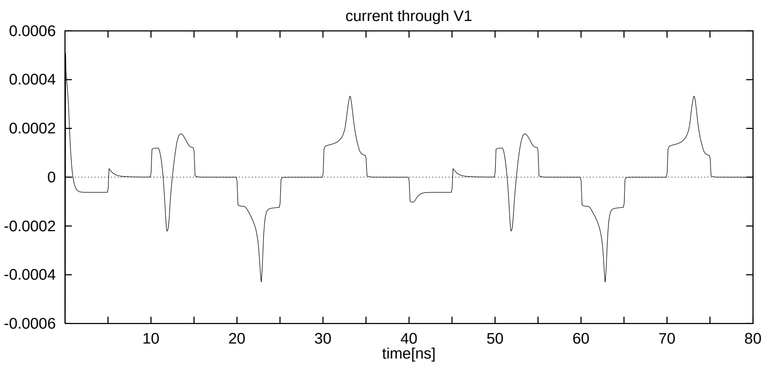


Fig. 5: Current $I_1$

In both cases, the current vanishes in the intervals [5ns, 10ns], [15ns, 20ns], [25ns, 30ns], [35ns, 40ns], [45ns, 50ns], [55ns, 60ns], [65ns, 70ns], and [75ns, 80ns]. In these intervals, both input signals $V_1$ and $V_2$ are constant.

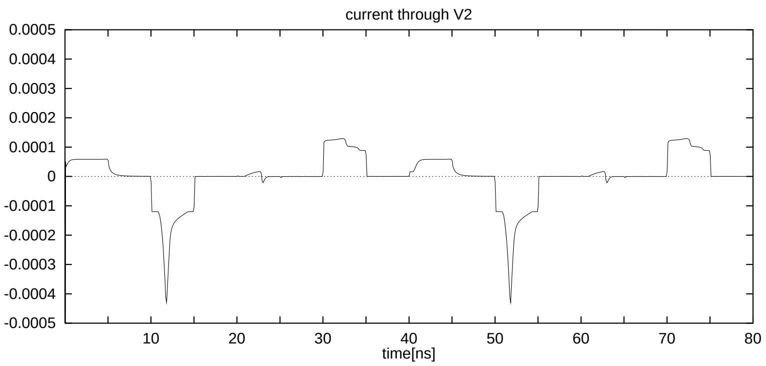


Fig. 6: Current $I_2$

All calculations were carried out by the BDF code DAE2SOL ([19]), which controlled order and stepsize via the smooth component

$$\tilde{P}\tilde{x}_\ell = \begin{pmatrix} P_A & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} q_\ell \\ x_\ell \end{pmatrix} = \begin{pmatrix} P_A q_\ell \\ 0 \end{pmatrix},$$

where

$$P_A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}.$$

From our experience, this control works essentially more effective than that of the complete value $\tilde{x}_\ell$.

As far as the weak instability term $\frac{1}{h_\ell} Q_{1\ell} H_\ell^{-1} \delta_\ell$ involved in the error estimation of Theorem 5.1 is concerned, this effect is even typical for index-2 DAEs. Besides the usual error propagation expected from the index 1 case, a certain defect component amplified by $h_\ell^{-1}$ influences the computation strongly.

In case of linear index 2 DAEs only the nullspace component of the solution $\tilde{Q}\tilde{x}_\ell$ is affected by that weak instability term (e.g. [13]). However, in nonlinear systems the situation is different. By means of a small academic example even in Hessenberg form, in [15] and [20] it is shown how weak instability may affect all solution components. A similar experience is reported in [1].

By the following table we realize those instability effects once more. The table shows the values computed by the constant stepsize backward Euler method with different stepsizes for approximating the currents $I_1(T) = 0$, $I_2(T) = 0$, $I_{DD}(T) = 0$, and $I_{BB}(T) = 0$, which have to vanish at the final point $T = 80 \cdot 10^{-9}$. The produced values reflect the theoretical results as expected. If we decrease the stepsize, the error becomes smaller up to the stepsize 2e-10. The error increases for stepsizes smaller than $2e - 10$. This clearly reflects the weak instability.

| stepsize | $I_1$ | $I_2$ | $I_{DD}$ | $I_{BB}$ |
|---|---|---|---|---|
| 8e-10 | 5.41e-10 | 1.25e-15 | 2.30e-09 | 3.29e-09 |
| 5e-10 | 2.36e-10 | 1.20e-15 | 1.00e-09 | 1.44e-09 |
| 2e-10 | 1.51e-10 | 1.19e-15 | 6.23e-10 | 9.00e-10 |
| 1e-10 | 1.88e-10 | 1.22e-15 | 4.91e-10 | 1.53e-10 |
| 5e-11 | 1.24e-09 | 2.25e-15 | 2.72e-09 | 5.34e-10 |

Sometimes it might be possible to handle the weak instability more effictively by improving the $\tilde{P}\tilde{Q}_{1,\ell}$-components of the approximations after each steps, or, which is in fact the same, by reducing the most dangerous parts of the defects, that is, those parts belonging to the range of $\tilde{Q}_{1,\ell}\tilde{G}_{2,\ell}^{-1}$ and of $Q_{1,\ell}H_\ell^{-1}$, respectively (cf. (4.9)).

More precisely, for a given approximation $\tilde{x}_\ell^0$ to $\tilde{x}_*(t_\ell)$ we try to determine the new approximation $\tilde{x}_\ell$,

$$\tilde{x}_\ell = (I - \tilde{P}\tilde{Q}_{1,\ell})\tilde{x}_\ell^0 + \tilde{P}\tilde{Q}_{1,\ell}\tilde{z}_\ell \tag{5.4}$$

by solving the equation

$$\tilde{Q}_{1,\ell}\tilde{G}_{2,\ell}^{-1}(\tilde{g}(\tilde{x}_\ell) - \tilde{r}(t_\ell)) = 0 \tag{5.5}$$

with respect to the correction term $\tilde{P}\tilde{Q}_{1,\ell}\tilde{z}_\ell$ we are looking for. This defect correction is nothing else but a generalization of the back-projection onto the right manifold, which was proposed by Ascher and Petzold ([2]) for Hessenberg systems.

THEOREM 5.2. *Under the conditions of Theorem 4.5, equation (5.5) is linear and it uniquely determines*

$$\tilde{P}\tilde{Q}_{1,\ell}\tilde{z}_\ell = \tilde{P}\tilde{Q}_{1,\ell}\tilde{x}_*(t_\ell).$$

*Proof.* The structural conditions lead to the two relations

$$\tilde{Q}_{1,\ell}\tilde{G}_{2,\ell}^{-1}(\tilde{g}(\tilde{x}_\ell) - \tilde{g}(\tilde{P}\tilde{x}_\ell)) = 0$$

and

$$\tilde{Q}_{1,\ell}\tilde{G}_{2,\ell}^{-1}(\tilde{g}(\tilde{x}_*(t_\ell)) - \tilde{g}(\tilde{P}\tilde{x}_*(t_\ell))) = 0.$$

With the notations given in the proof of Theorem 4.5 we have

$$\tilde{Q}_{1,\ell}\tilde{G}_{2,\ell}^{-1} = \begin{pmatrix} P_A g_\ell' Q_{1,\ell} H_\ell^{-1} & P_A g_\ell' Q_{1,\ell} R_{*,\ell} \\ Q Q_{1,\ell} H_\ell^{-1} & Q Q_{1,\ell} R_{*,\ell} \end{pmatrix}.$$

Putting $\tilde{z}_\ell = \begin{pmatrix} \gamma_\ell \\ z_\ell \end{pmatrix}$ we obtain

$$\tilde{P}\tilde{Q}_{1,\ell}\tilde{z}_\ell = \begin{pmatrix} P_A g_\ell' Q_{1,\ell} R_{*,\ell}\gamma_\ell \\ 0 \end{pmatrix},$$

further, from (5.5),

$$Q_{1,\ell}R_{*,\ell}\gamma_\ell = Q_{1,\ell}R_{*,\ell}g(0) - Q_{1,\ell}H_\ell^{-1}(f(0) - r(t_\ell)). \tag{5.6}$$

On the other hand, the relation $\tilde{Q}_{1,\ell}\tilde{G}_{2,\ell}^{-1}(\tilde{g}(x_*(t_\ell)) - \tilde{r}(t_\ell)) = 0$ given for the exact DAE solution leads to

$$Q_{1,\ell}R_{*,\ell}q_*(t_\ell) = Q_{1,\ell}R_{*,\ell}g(0) - Q_{1,\ell}H_\ell^{-1}(f(0) - r(t_\ell)).$$

□

REMARK: In practical computations we do not have the exact projectors $\tilde{Q}_{1,\ell} = \tilde{Q}_1(x_*(t_\ell))$ etc., but instead we have to use the approximations $\tilde{Q}_1(x_\ell^0)$. If im $\tilde{Q}_1(y)$ does not rotate too quickly and if $\tilde{x}_\ell^0$ is close enough to $x_*(t_\ell)$, we may expect the defect correction to work well.

**6. Final Remark.** For the classical formulation (2.1), one can think of defining index-2 tractability also for nullspaces $N(x)$ rotating with $x$. But, again we should expect a higher perturbation index and consequently much harder numerical difficulties. On the other hand, the BDF is known to fail for index-2 systems with rotating leading nullspaces, even for very simple systems. It is known that the exponential numerical instability is the reason for that.

Hence, there is only some hope to handle more general index-2 systems (2.1) which do not satisfy (3.3) by means of the BDF if the nullspace remains somewhat restricted to certain index-1 parts of the system. Possibly, then only weak instabilities will arise. However, this question needs a further great theoretical effort as well as a very deep insight into the circuit structure.

It should be stressed once more that the charge/flux oriented formulation (2.2)-(2.3) stands out for its *constant* leading nullspace.

**Acknowledgements.** The authors gratefully acknowledge the referees for their careful reading of the manuscript and their valuable suggestions to improve the readability of this paper.

REFERENCES

[1] M. Arnold, *Applying BDF to quasilinear differential-algebraic equations of index 2: perturbation analysis.* in preparation, 1995.

[2] U. Ascher and L. Petzold, *Stability of computational methods for constrained dynamics systems*, SIAM J. Sci. Stat. Comput., (1991), pp. 95–120.

[3] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *The Numerical Solution of Initial Value Problems in Ordinary Differential-Algebraic Equations*, North Holland Publishing Co., 1989.

[4] R. Bulirsch and A. Gilg, *Effiziente numerische Verfahren für die Simulation elektrischer Schaltungen*, in Informatik in der Praxis, H. Schwärtzel, ed., Springer Verlag, 1986, pp. 3–12.

[5] G. Denk and P. Rentrop, *Mathematical models in electric circuit simulation and their numerical treatment*, in Teubner-Texte zur Mathematik, vol. 121, 1991, pp. 305–316.

[6] U. Feldmann, 1995. private communication.

[7] U. Feldmann and M. Günther, *The dae-index in electric circuit simulation*, in Proc. IMACS Symposium on Mathematical Modelling, I. Troch and F. Breitenecker, eds., 4, 1994, pp. 695–702.

[8] U. Feldmann, U. A. Wever, Q. Zheng, R. Schultz, and H. Wriedt, *Algorithm for modern circuit simulation*, Archiv für Elektronik und Übertragungstechnik, 46 (1992), pp. 274–285.

[9] E. Griepentrog and R. März, *Differential-Algebraic Equations and Their Numerical Treatment*, Teubner-Texte zur Mathematik No. 88, BSB B.G. Teubner Verlagsgesellschaft, Leipzig, 1986.

[10] M. Günther and U. Feldmann, *CAD modelling of electric circuits: A classification with respect to structure and index.* in preparation, 1995.

[11] M. Günther and P. Rentrop, *Suitable one-step methods for quasilinear-implicit odes*, Tech. Report TUM-M9405, Mathematisches Institut, TU München, 1994.

[12] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II: Stiff and differential-algebraic problems*, Springer Series in Computational Mathematics 14, Springer-Verlag, Berlin, Heidelberg, 1991.

[13] R. März, *Numerical methods for differential-algebraic equations*, Acta Numerica, (1992), pp. 141–198.

[14] ———, *On linear differential-algebraic equations and linearizations*, APNUM, 18 (1995), pp. 267–292.

[15] R. März and C. Tischendorf, *Solving more general index 2 differential algebraic equations*, Comp. and Math. with Appl., 28 (1994), pp. 77–105.

[16] W. Mathis, *Theorie nichtlinearer Netzwerke*, Springer Verlag Berlin Heidelberg NewYork, 1987.

[17] P. Rabier and W. Rheinboldt, *A general existence and uniqueness theory for implicit differential-algebraic equations*, Differential and Integral Equations, 4 (1991), pp. 563–582.
[18] H. Shichman and D. A. Hodges, *Insulated-gate field-effect transistor switching circuits*, IEEE J. Solid State Circuits, SC-3 (1968), pp. 285–289.
[19] C. Tischendorf, *Die BDF für nichtlineare Algebro-Differentialgleichungen vom Index 2*, 1992. Diplomarbeit.
[20] ———, *Feasibility and stability behaviour of the BDF applied to index-2 differential algebraic equations*, ZAMM, 75 (1995), pp. 927–946.
[21] ———, *Solution of index-2 differential algebraic equations and its application in circuit simulation.* in preparation, 1995.
[22] R. Winkler, *On simple impasse points and their numerical computations*, Tech. Report 94-15, Fachbereich Mathematik, Humboldt-Univ. zu Berlin, 1994.